Articulation in time

Some word-initial segments in Swedish

Svensson Lundmark, Malin

2020

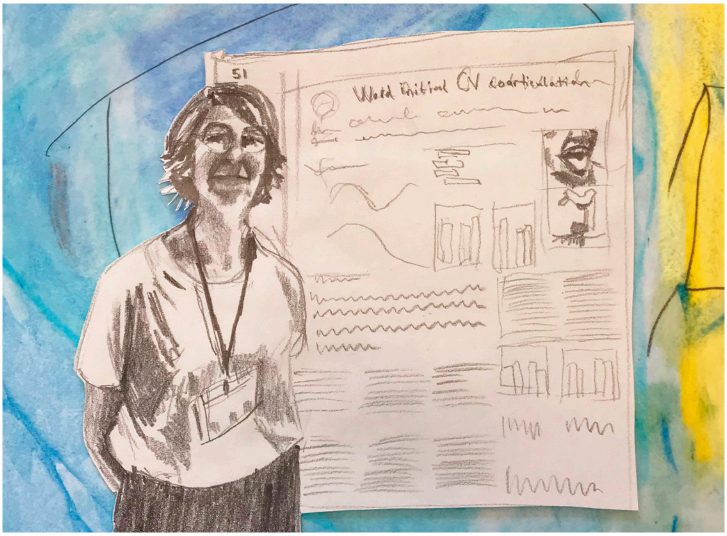
LUND UNIVERSITY

PO Box 117
221 00 Lund
+46 46-222 00 00

Articulation in time

Some word-initial segments in Swedish

MALIN SVENSSON LUNDMARK
CENTRE FOR LANGUAGES AND LITERATURE | LUND UNIVERSITY



Articulation in time

Some word-initial segments in Swedish

Speech is both dynamic and distinctive at the same time. This implies a certain contradiction which has entertained researchers in phonetics and phonology for decades. The present dissertation assumes that articulation behaves as a function of time, and that we can find phonological structures in the dynamical systems. EMA is used to measure mechanical movements in Swedish speakers. The results show that tonal context affects articulatory coordination. Acceleration divides the movements of the jaw and lips into intervals of postures and active movements. These intervals are affected differently by the tonal context. Furthermore, a bilabial consonant is shorter if the next consonant is also made with the lips. A hypothesis of a correlation between acoustic segment duration and acceleration is presented. The dissertation highlights the importance of time for how speech ultimately sounds. Particularly significant is the combination of articulatory timing and articulatory duration.

Joint Faculty of Humanities and Theology
Centre for Languages and Literature

ISBN 978-91-89213-20-3

Printed by Media-Tryck, Lund 2020 NORDIC SWAN ECOLABEL 3041 0903

Articulation in time

Some word-initial segments in Swedish

Malin Svensson Lundmark

DOCTORAL DISSERTATION
by due permission of the Joint Faculty of Humanities and Theology,
Lund University, Sweden.
To be defended at LUX:C121, Helgonavägen 3, Lund
Friday 23 October, 2020, 16:00

Faculty opponent
Professor Donna Erickson

Haskins Laboratories, New Haven, CT, USA

Organization
LUND UNIVERSITY

Centre for Languages and Literature

Document name
DOCTORAL DISSERTATION

Date of issue: 23 October 2020

Author: Malin Svensson Lundmark

Sponsoring organization

Title and subtitle
Articulation in time. Some word-initial segments in Swedish.

Abstract

The present dissertation is a contribution to speech production modelling. It assumes that speech is dynamic, and that articulation behaves as a function of time. Certain aspects of both timing and duration appear to be very important components of a phonological unit. However, many phonologies do not take time into account. This might be because it is not often assumed that dynamical systems govern both us and our language. This dissertation focuses on measuring the mechanical movements during articulation, in order to enhance our understanding of which the phonological units are. Also, Swedish word accent is examined.

Movements that, under the effect of a variable, are systematic over several speakers are considered more likely to form part of a phonology than those which vary more both between and within speakers. For this purpose, articulatory movements of a total of 23 speakers have been recorded with ElectroMagnetic Articulography. The data analyses show: the way we measure the start of a movement affects whether it can be considered to have good timing or not (Paper 2); the creation of tones is integrated with the creation of consonants and vowels (Paper 1); tonal context affects movements of lips, jaw and tongue body (Paper 2 and 3); a long vowel is executed through a longer open jaw (Paper 3); the acceleration profile of the jaw has clear systematic features (Paper 3); when any of the consonants in a CVC sequence is made with the same articulator as the other consonant, both segments are shortened, regardless of place of articulation (Paper 4).

On top of this, the dissertation contains a more detailed introductory chapter in which the following hypotheses are emphasized: 1) the articulators responsible for the f_0 rise and f_0 fall, respectively, may be timed with other active articulators. This timing, which is assumed to be phonological, appears to have a biomechanical effect on either articulator; 2) bilabial and jaw movements seem to consist of active intervals and postures defined on the basis of maximal acceleration and deceleration; 3) acoustic segment duration seems to coincide with the acceleration profile of the consonant articulation.

All in all, the dissertation demands continued work with both mechanical measurements and dynamic model developments. Mapping the systems of motion that we already know exist helps us to understand the function of time in speech and to reveal the real phonological units.

Key words
Acceleration, articulation, articulography, coarticulation, EMA, gesture, phonology, prosody, speech motor control, speech production modelling, Swedish, word accent, tone

Classification system and/or index terms (if any)

Supplementary bibliographical information

Language: English

ISSN and key title

ISBN: 978-91-89213-20-3 (print)

Recipient’s notes

Number of pages: 224

Price

Security classification

I, the undersigned, being the copyright owner of the abstract of the above-mentioned dissertation, hereby grant to all reference sources permission to publish and disseminate the abstract of the above-mentioned dissertation.

Signature

Date 2020-09-24

Articulation in time

Some word-initial segments in Swedish

Malin Svensson Lundmark

LUND
UNIVERSITY

Cover art by Malin Svensson Lundmark *Still life with flowers, orange and hyoid bone* and *Self-portrait Lisbon*

Joint Faculty of Humanities and Theology
Centre for Languages and Literature

ISBN 978-91-89213-20-3 (print)
ISBN 978-91-89213-21-0 (digital)

Printed in Sweden by Media-Tryck, Lund University
Lund 2020

NORDIC SWAN ECOLABEL
Printed matter
3041 0903

Media-Tryck is a Nordic Swan Ecolabel
certified provider of printed material.
Read more about our environmental
work at www.mediatryck.lu.se

MADE IN SWEDEN

To my grandmother Anna-Lisa Andersson

Don't you wonder sometimes
'bout sound and vision

David Bowie

Table of Contents

Abstract **11**

Acknowledgments **13**

Publications and contributors **17**

Non-included papers and contributions..18

List of abbreviations **21**

1 Introduction..**23**

1.1 Background..24

1.2 Research questions and Theoretical background.................................25

1.2.1 Questions on communicative efficiency 26

1.2.2 Questions on the dynamic nature of speech 27

1.2.3 Questions on Swedish phonology..34

1.3 Limitations of previous studies 39

1.4 Scope of the dissertation..40

1.5 Outline of the dissertation...41

2 Methods..**43**

2.1 Speech material 43

2.1.1 The pilot study..43

2.1.2 The corpus 44

2.2 Procedures..48

2.2.1 Articulography 48

2.2.2 Preprocessing 51

2.2.3 Recordings 51

2.2.4 Post-processing...54

2.3 Measurements 56

2.3.1 Paper 1 – Acoustic and articulatory measurements 57

2.3.2 Paper 2 – Articulatory and acoustic measurements 59

2.3.3 Paper 3 – Articulatory measurements 62

2.3.4 Paper 4 – Acoustic measurements..64

2.4 Statistics and analysis 65

3 The studies 67

3.1 Paper 1: Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents 67

3.1.1 Summary 67

3.1.2 General discussion 68

3.2 Paper 2: Word-initial consonant-vowel coordination in a lexical pitch-accent language 69

3.2.1 Summary 69

3.2.2 General discussion 70

3.3 Paper 3: Jaw movements in two tonal contexts 71

3.3.1 Summary 71

3.3.2 General discussion 72

3.4 Paper 4: Mutual influence of word-initial and word-medial consonantal articulation 73

3.4.1 Summary 73

3.4.2 General discussion 75

4 Discussion 77

4.1 Cohesion between articulators 77

4.1.1 How does tone affect inter-articulator cohesion? 79

4.1.2 Implications for the Swedish word accents 80

4.2 Linking articulation with acoustics I 81

4.2.1 Locus equation as an explanatory model for articulation 82

4.2.2 Formant changes that lead to a new way of thinking 84

4.3 Evaluation of some selected measurements 85

4.3.1 CV time lags 86

4.3.2 Acceleration-based intervals 87

4.3.3 Peak velocity 91

4.4 Linking articulation with acoustics II 93

4.4.1 Articulation in time 93

4.4.2 Towards a theory of time and segment boundary 94

4.4.3 A time-locked acceleration-to-acoustic-boundary hypothesis of place of articulation 96

4.5 Communicative efficiency 98

4.5.1 An “immune” and contrastive articulation 99

4.5.2 On hard and easy words 100

4.6 Summary conclusions 101

5 Future research 103

5.1 The significance of articulation in time for modelling speech production 104

5.1.1 Towards an articulatory hierarchy 105

5.1.2 Explaining acoustic segment phenomena 107

5.2 The Swedish word accents 108

5.3 Other implications and unresolved issues 109

5.3.1 Speech motor control 109

5.3.2 Understanding speech disorders 110

5.3.3 What do the results mean for future perceptual research? 111

Populärvetenskaplig sammanfattning på svenska 113

References 115

Appendices 123

Attached papers 127

Abstract

The present dissertation is a contribution to speech production modelling. It assumes that speech is dynamic, and that articulation behaves as a function of time. Certain aspects of both timing and duration appear to be very important components of a phonological unit. However, many phonologies do not take time into account. This might be because it is not often assumed that dynamical systems govern both us and our language. This dissertation focuses on measuring the mechanical movements during articulation, in order to enhance our understanding of which the phonological units are. Also, Swedish word accent is examined.

Movements that, under the effect of a variable, are systematic over several speakers are considered more likely to form part of a phonology than those which vary more both between and within speakers. For this purpose, articulatory movements of a total of 23 speakers have been recorded with ElectroMagnetic Articulography. The data analyses show: the way we measure the start of a movement affects whether it can be considered to have good timing or not (Paper 2); the creation of tones is integrated with the creation of consonants and vowels (Paper 1); tonal context affects movements of lips, jaw and tongue body (Paper 2 and 3); a long vowel is executed through a longer open jaw (Paper 3); the acceleration profile of the jaw has clear systematic features (Paper 3); when any of the consonants in a CVC sequence is made with the same articulator as the other consonant, both segments are shortened, regardless of place of articulation (Paper 4).

On top of this, the dissertation contains a more detailed introductory chapter in which the following hypotheses are emphasized: 1) the articulators responsible for the f_0 rise and f_0 fall, respectively, may be timed with other active articulators. This timing, which is assumed to be phonological, appears to have a biomechanical effect on either articulator; 2) bilabial and jaw movements seem to consist of active intervals and postures defined on the basis of maximal acceleration and deceleration; 3) acoustic segment duration seems to coincide with the acceleration profile of the consonant articulation.

All in all, the dissertation demands continued work with both mechanical measurements and dynamic model developments. Mapping the systems of motion that we already know exist helps us to understand the function of time in speech and to reveal the real phonological units.

Acknowledgments

During the years I have trained to become a researcher, some special people have shared their knowledge with me in the best way possible. First, thanks to my supervisor Susanne Schötz who put me on an early pilot recording and has taught me everything in the lab. Without that kickstart and her dedication, this work would not have been able to accelerate. I feel great gratitude to my supervisor Sven Strömqvist who took over the main supervisor responsibility in the very best way. Your understanding, sensitivity, perspective, and strategic eye helped me sew everything together, and it has been very satisfying to talk to you about the writing. You may already know this, but the work sort of became so much easier when you came into the picture. Thank you for all your encouragement and kind words that have made the journey towards finding my path easier! I also want to thank my unofficial supervisor Johan Frid for assisting me with quick solutions to logistic and mathematical problems. You have a natural talent for a dynamic way of thinking and I really appreciate our cooperation and your interest in the issues at hand. Thanks to my unofficial supervisor Martine Grice who for just a few important months, step by step, and with a warm hand, helped me regain my self-esteem as a researcher. And last, but definitely not least, a huge thank you to my constant co-supervisor Gilbert Ambrazaitis for all the conversations we have had throughout the years, for your involvement in my process, your sometimes stubborn desire to understand, and for your friendship. You have been a great support to me and your confidence in my skills gave me strength. I can't thank you enough. I feel extremely grateful to both Sven and Gilbert for your work "behind the scenes" during the final phase.

I'm honoured to have Donna Erickson as faculty opponent for the defense of my dissertation, and to have Doris Mücke, Philip Hoole and Mikael Roll as my academic committee. Thanks to Anders Löfqvist, Birgitta Sahlén and Joost van de Weijer for acting as reserves, and to Johannes Persson for acting as chair. Thank you to mock opponent Mattias Heldner for challenging questions and helpful comments. Thanks to Lars-Håkan, Christina, David and others for help with proofreading.

Many thanks to current and former employees of the Linguistics department at the Centre of Languages and Literature, Lund University. As I like to say: it takes a village to raise a doctoral student. Thank you for doing your very best, and for encouraging me along the way. Special thanks to the neurolinguistics group who challenged and inspired me to find the right direction in my research. A special thank you also to my former and current doctoral colleagues who, like siblings, have supported me by just being there. Thanks to all those people at Lund University who have been service-minded and helped me with various questions, maybe especially in the spring of 2019. You know who you are - thank you! A huge thank you also to the Lund University Humanities Lab. It goes without saying that my dissertation could never have been written had it not been for these fantastic facilities. The same applies to the participants. If you had not volunteered, whether out of helpfulness or curiosity, the work would not have been possible. Thank you so much!

A big thank you to all the researchers who gave their time to talk to me at conferences and otherwise. These conversations, big and small, have meant a lot to me. Some who gave a little extra of their time I would like to thank in particular (in alphabetical order): Anders Löfqvist, Anne Cutler, Aude Noiray, Bettina Braun, Briana van Epps, Christine Mooshammer, David House, Doris Mücke, Eva Åkesson, Frida Blomberg, Joost van de Weijer, Marianne Gullberg, Mechtild Tronnier, Merle Horne, Niclas Burenhult, Oxana Rasskazova. An extra big thank you to Man Gao for helpfulness and generosity in everything related to articulatory phonological thinking. Special thanks for the hospitality and the many good conversations I had in Cologne during my visit there. It was short but sweet. Thanks also to all the phonetic people at various conferences and seminars, who were accommodating and listened to my rather tentative ideas. Thanks also to all the anonymous reviewers who have devoted time and energy to my texts. Your comments have been most valuable during this trip. Thank you, whoever you are.

I also want to take this opportunity to thank all the teachers I have had since I was a little girl, who have seen me, believed in me and encouraged me: Eva, Jan-Anders, Monica, Ingvar, Alf, Thomas, Gösta and others. Now school is over. School's out forever.

Thanks to dear friends who, in some strange way, always manage to make me enjoy myself and my life. Special thanks to Sara who sometimes gives me shelter. Thanks to my mom and dad for giving me this world. Thank you for giving me perseverance and naivety, as well as an unwavering sense of right and wrong about most things. Thank you also for giving me my dear sisters. Anna, Cilla, Lina - you are the best sisters, I argue, a woman can have, and I love you very much. But perhaps the biggest thank you goes to the small family inside the big family. First, loving thank you to Fabian who shares almost everything with me. Words are not enough or I am not yet capable of expressing my love in words. Let me say this, I look forward to spending the next thirty years or so with you. My children Nils, Julie and Lo, you are what motivates and gives meaning to me. You are my joy and my inspiration. How can I even begin to describe my love for you. Thanks to all the wonderful caregivers and educators who gave their time to my children so that I in turn have been able to learn how to become a researcher.

This book is dedicated to my grandmother Anna-Lisa Andersson, who was a housewife all her life. Through her curiosity about the visual arts, about music and poetry, she unknowingly taught me how to analyze and make connections (as proof of that: notes and embedded clippings in countless books), and how enjoyable such work can be. In many ways, she was the first researcher I came into contact with.

Finally, to my daughter Julie, I want to say: Nu är boken färdig!

Publications and contributors

The studies in this dissertation have been carried out in collaboration with others. The details of these collaborations are given below.

Paper 1 Svensson Lundmark, M., Ambrazaitis, G., & Ewald, O. (2017). Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents. In *Proceedings of Interspeech 2017,* Stockholm, Sweden, 3236–3240.

Gilbert Ambrazaitis helped with the planning and the design of the study, read and commented on the manuscript, and together with Otto Ewald helped with acoustic analyses.

Paper 2 Svensson Lundmark, M., Ambrazaitis, G., Frid, J., & Schötz, S. (submitted). Word-initial consonant-vowel coordination in a lexical pitch-accent language.

Gilbert Ambrazaitis helped with the planning of the study, acoustic analysis, and together with Susanne Schötz and Johan Frid with reading and commenting on the manuscript. In addition, Johan Frid helped with acoustic and articulatory analyses, and together with Susanne Schötz with setting up the data collection.

Paper 3 Svensson Lundmark, M., & Frid, J. (2019). Jaw movements in two tonal contexts. In *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia, 1843–1847.

Johan Frid helped with articulatory analyses and together with Gilbert Ambrazaitis with reading and commenting on the manuscript.

Paper 4 Svensson Lundmark, M. (manuscript). Mutual influence of word-initial and word-medial consonantal articulation.

Martine Grice helped with the design of the study as well as with reading and commenting on first drafts. Mattias Heldner, Sven Strömqvist, Gilbert Ambrazaitis and Johan Frid contributed with reading and commenting on the manuscript.

Non-included papers and contributions

Below is a list of studies that have been presented at conferences or published in conference proceedings, but did not in the end become part of the dissertation. Some are directly related to, and in some respects absolutely crucial to, the work presented in the dissertation. Others, which are related to side projects and were carried out during my time as a doctoral student, are also listed.

Studies related to the dissertation

Svensson Lundmark, M., Frid, J., & Schötz, S. (2015). A pilot study: acoustic and articulatory data on tonal alignment in Swedish word accents. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK, Paper number 590.

Svensson Lundmark, M. (2017). Coordination of Word Onset Articulatory Gestures in Swedish: Anticipatory Cues to Word Accents. *Stem-, Spraak- en Taalpathologie*, *22* (Supplement).

Svensson Lundmark, M. (2017). *Intra-syllabic structures of articulatory gestures in Swedish prosody* [Paper presentation]. 3rd Doctorial Consortium, Stockholm, Sweden.

Svensson Lundmark, M. (2018). Durational properties of word-initial consonants – an acoustic and articulatory study of intra-syllabic relations in a pitch-accent language. In *Proceedings of Fonetik 2018,* Gothenburg, Sweden, 65–66.

Svensson Lundmark, M., & Frid, J. (2018). *Word onset CV coarticulation affected by post-vocalic consonants*. Poster session presented at LabPhon16, Lisbon, Portugal.

Svensson Lundmark, M., Frid, J., Ambrazaitis, G., & Schötz, S. (2018a). *Word-initial CV coarticulation in a pitch-accent language.* Abstract from International Conference on Tone and Intonation TIE2018, Gothenburg, Sweden.

Svensson Lundmark, M., Frid, J., Ambrazaitis, G., & Schötz, S. (2018b). *The effect of Swedish Word Accent on word initial CV coarticulation*. Abstract from Phonology in the Nordic Countries (FiNo) 2018, Lund, Sweden.

Studies related to side projects

Ambrazaitis, G., Svensson Lundmark, M., & House, D. (2015a). Head beats and eyebrow movements as a function of phonological prominence levels and word accents in Stockholm Swedish news broadcasts. Abstract from 3rd European

Symposium on Multimodal Communication (MMSYM 2015), Dublin, Ireland.

Ambrazaitis, G., Svensson Lundmark, M., & House, D. (2015b). Head Movements, Eyebrows, and Phonological Prosodic Prominence Levels in Stockholm Swedish News Broadcasts. FAAVSP - The 1st Joint Conference on Facial Analysis, Animation, and Auditory-Visual Speech Processing.

Ambrazaitis, G., Svensson Lundmark, M., & House, D. (2015c). Multimodal levels of prominence: a preliminary analysis of head and eyebrow movements in Swedish news broadcasts. In *Proceedings of Fonetik 2018,* Gothenburg, Sweden, 11–16.

Frid, J., Ambrazaitis, G., Svensson Lundmark, M., & House, D. (2016). Towards classification of head movements in audiovisual recordings of read news. Abstract from 4th European and 7th Nordic Symposium on Multimodal Communication (MMSYM 2016), Copenhagen, Denmark.

Gao, M., Svensson Lundmark, M., Schötz, S., & Frid, J. (2018). A Cross-Language Study of Tonal Alignment in Scania Swedish and Mandarin Chinese. 12. Poster session presented at Phonology in the Nordic Countries (FiNo) 2018, Lund, Sweden.

Frid, J., Gao, M., Svensson Lundmark, M., & Schötz, S. (2018). *Pitch-to-segment Alignment in South Swedish and Mandarin Chinese: A Cross-language Comparison.* Abstract from International Conference on Tone and Intonation TIE2018, Gothenburg, Sweden.

Frid, J., Svensson Lundmark, M., Ambrazaitis, G., Schötz, S., & House, D. (2018). EMA-based head movements and phrasing: a preliminary study. In *Proceedings of Fonetik 2018,* Gothenburg, Sweden, 17–20.

Frid, J., Svensson Lundmark, M., Ambrazaitis, G., Schötz, S., & House, D. (2019) Investigating visual prosody using articulography. In *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries: Copenhagen, March 6-8 2019* CEUR.

List of abbreviations

A1	Swedish word accent 1
A2	Swedish word accent 2
AG501	An electromagnetic articulograph by the Carstens company
AIC	Akaike Information Criterion
AP	Articulatory Phonology
C	Consonant
C:	Long consonant
C1	First consonant of the word
C2	Second consonant of the word
C3	Third consonant of the word
CNS	Central Nervous System
CT	Cricothyroid muscle
CV	Consonant-Vowel sequence
DST	Dynamical Systems Theory
EL	Left Ear
EMA	ElectroMagnetic Articulograph
f_0	Fundamental frequency of oscillation of the vocal folds (also referred to as f0 or F0 in the papers)
F_2	Second formant frequency (also referred to as F2 in the papers)
F_3	Third formant frequency (also referred to as F3 in the papers)
GLMMs	Generalized Linear Mixed effects regression Models
H	High phonological tone
NR	Nose Ridge

JW	Jaw
L	Low phonological tone
LA	Lip Aperture
LE	Locus Equation
LL	Lower Lip
ms	milliseconds
SWA/SWAs	Swedish Word Accent/s
TB	Tongue Body (or Tongue Blade, when specified)
TB1	Corresponding to tongue blade
TB2	Corresponding to tongue dorsum
TD	Tongue Dorsum
TD model	Task Dynamics model
TT	Tongue Tip
UL	Upper Lip
VCV	Consonant-Vowel-Consonant sequence
V	Vowel
V:	Long vowel
V1	First vowel of the word
V2	Second vowel of the word
VC	Vowel-Consonant sequence

1 Introduction

Human language sounds the way it does because the speech apparatus is what it is. Thus, acoustics is dependent on articulation. With a little imagination, the relationship between articulation and acoustics can be described as the relationship between a footstep and its footprint. The mass and the velocity of the foot (and of the entire foot carrier) determine the final result of the footprint. If the foot carrier is moving fast, the footprint reflects this. If the foot carrier is a small or a large individual, this can be discerned by means of the depth and the length of the footprint. In addition, the recipient of the message (the listener) has primarily access only to the footprint, and this helps her/him interpret what the foot carrier (the speaker) communicates. Some things may be easier to interpret than others, for example if the foot carrier is small or large, but the nuances of the movements are perhaps a little trickier. Either way, it is necessary for the recipient of the message to interpret the footprint according to her expectations and acquired knowledge about how a foot moves, that is, the natural movement pattern of the body and the foot. Nothing else can be of equal importance.

In this dissertation, I argue that we find a similar relationship between articulation and acoustics – and, by extension, perception. That is, the listener (the recipient) hears the acoustics (sees the footprint), but necessarily interprets it based on motor knowledge and prior experience of the movements of the speech apparatus (the footstep) of the speaker (the foot carrier). In other words, the acoustic signal contains cues and clues as how the articulators move, their speed and their mass. The aim of this dissertation is to contribute to our knowledge of how systematic articulatory movements are involved in speech modelling.

During the time I have been working on this dissertation, a clear theme has emerged that is quite easy to describe, despite the, sometimes, complex nature of the subject. Quite simply, all four studies in this dissertation are about consonants and vowels, and how they differ in their most basic parts. Thus, what initially might appear to the reader to be another chapter in the history of research on Swedish word accent is really about how consonants and vowels are affected by various intra-syllabic parameters, of which tones, for example, are one.

But, why, you may ask, is the difference between consonants and vowels significant? After all, they only follow upon one another: consonant, vowel, consonant, vowel, etc., like beads on a string. Unfortunately, that is a simplified picture, and a shortcut that can lead to a dead end. Historically, phonetics is based on the three cornerstones *articulation*, *acoustics* and *perception*. Much work has been done within each field, but it is important, if not decisive, for the different research orientations themselves, how these three cornerstones interconnect. Perception research, for example, is based on acoustic research and certain assumptions about the link between articulation and acoustics. So not to assume that consonants and vowels are actually motorically overlapping and physically coordinated, which has a major impact on their acoustic patterns, can therefore have fatal results for, for example, a listening test. Therefore, developing research on how articulation and acoustics are specifically related to each other may not only benefit these two fields of research, but also have a significant impact on perception research, not to mention other branches of language research.

Thus, if we return to the foot, how can the receiver, based on her knowledge of how the foot can move, get all the necessary information out of a footprint? There must be a system throughout the physiological meeting between the foot and the substrate that allows this. A system that also applies to speech; the various articulators and their movements in the oral cavity. Thus, a structured system is needed that is distinctive and at the same time open to the dynamics of movement. Quite simply, what is needed may be a phonology based on the laws of physics.

1.1 Background

As a research topic, phonetics is a hybrid of the humanities and the natural sciences. At its core, phonetics is thus an interdisciplinary research topic, and our collaborations occur naturally across the spectrum. Common to all phonetics is a holistic approach to human language. We thus naturally assume the limitations and possibilities of the human body, and always on the basis of a linguistic question. This may have a particular bearing on the research topic with which this dissertation is concerned, since one measures the body's movements in the search for phonological distinction.

The link between articulation and acoustics is however not a new field of research. On the contrary, phoneticians have at all times been looking for the systematic articulatory movements that underlie the distinctive phonological units, a search often guided by acoustic patterns. An important milestone that may be mentioned here is the knowledge of how consonants overlap with vowels. Like underlying diphthongs, vowels are combined in one layer, while consonants act as islands on top of that layer. This connection was demonstrated by Sven Öhman as early as 1966, through an acoustic study of VCV sequences. Öhman demonstrated very clearly that the overlapping constriction, that is the consonant, was firmly in line with the start of the second vowel movement, coupled like syllables, one could say: V-CV. Thus, in other words, the co-articulation is greater in a CV sequence than a VC sequence, a conclusion shared by many others since then (MacNeilage & DeClerk, 1969; Browman & Goldstein, 1988; Fowler & Saltzman, 1993; Byrd, 1995; Löfqvist & Gracco, 1999; Recasens, 2002). With the help of technological advances, our knowledge of articulatory mobility has progressed. We have been informed, for example, that the lips, and the jaw, are correlated in time by their highest velocity (Gracco, 1988), or that the jaw is more or less open depending on which type of consonant the tongue tip is making (Lindblom, 1983; Mooshammer et al., 2006).

A large part of the work presented in this dissertation work is of practical nature and has been carried out with an apparatus and a particular method for measuring articulatory movement in time and space, the ElectroMagnetic Articulograph (EMA). The nature of this machine is reviewed in the method chapter of this dissertation. However, its potential has given rise to increasing use over the last thirty years, as seen through several studies. For example, its possibilities for inter-articulatory timing measurements has made EMA suitable for speech modelling in linguistic research (Löfqvist & Gracco, 1999, 2002; Recasens, 2002; Löfqvist, 2007; Gao, 2008; Hoole et al., 2009; Mücke et al., 2012; Stone, 2013; Erickson et al., 2014; Tilsen, 2016; Shaw & Chen, 2019; Pastätter & Pouplier, 2017). In particular, the high resolution of time appears appropriate for studies on co-articulation. Prosodic research, of which this dissertation is part, has also gained momentum through the use of its methodology, by examining the articulatory constraints of phrase boundary (Cho, 2002; Byrd et al., 2005; Hoole et al., 2009; Mooshammer et al., 2013; Bombien et al., 2013; Erickson et al., 2014), and accents and tones (Cho, 2002; Erickson et al., 2004; D'Imperio et al., 2007; Gao, 2008; Mücke et al., 2012; Yi & Tilsen, 2014; Mücke & Grice, 2014; Niemann et al., 2014; Katsika et al., 2014; Shaw et al., 2016). Articulatory measurements with focus more on spatial position have also been made using the EMA (Erickson et al., 2004; Mooshammer et al., 2007; Shaw et al., 2016), including some on Swedish speakers (Schötz et al., 2013). Articulography also allows the recording of facial movements other than those of the articulators, which has been shown in studies on multimodality (Krivokapic et al., 2017; Frid et al., 2019).

1.2 Research questions and Theoretical background

The following sections introduce the theoretical framework used in this dissertation. First, a discussion of articulatory effort versus perceptual contrast is presented. This part serves as a declaration of why word-initial position is a main thread in the dissertation. After that follows an exposition of dynamic speech, how it behaves and what is believed to be its structure. The theories and models mentioned here may at first be seen as separate, but they are really based on each other, as ramifications or levels of the same theoretical framework. Therefore, the emphasis is on the issues they have in common, although they are presented separately in the next section for the sake of simplicity. The theoretical background concludes with an introductory text on Swedish phonology, focusing on Swedish word accent, since this is the topic of three of the four dissertation studies.

1.2.1 Questions on communicative efficiency

It is usually assumed that the phonological inventories of different languages are based on two rules: perceptual contrast should be high, and articulatory effort low (Lindblom & Maddieson, 1988). Furthermore, it is usually assumed that a word-initial position is most important for the listener due to incremental processing; and that signals are gradually made available to the listener over time (Marslen-Wilson & Zwitserlood, 1989; Cutler, 2012; Beddor et al., 2013). This means that, in a word-initial position, the perceptual contrast should be as high as possible, while the articulatory effort should be as low and yet as effective as possible. In a word-medial position, equally strong requirements would not apply. In fact, phonetic information in segments seems to depend on placement: listeners have an easier time identifying word-initial segments than word-medial segments (for a review on word processing, see Wedel et al., 2019). This supports the Lindblom/Maddieson assumption of articulatory effort and perceptual contrast.

Because word-initial position is significant, phonological rules also seem to aim for as a high lexical contrast as possible at the beginning of words (Wedel et al., 2019). Another aspect of communicative efficiency, and highly relevant to the above, is that more frequent words tend to be shorter, while less frequent words are longer (Zipf's law of abbreviation; for a review, see Wedel et al., 2019). Long words seem to have to contribute more lexical information, because the context does not, as if less frequent words need to be more specified. However, short and long words can also be more or less predictable. In this regard, phonetic information seems to play a greater role for the listener (Wedel et al., 2019).

To summarize, it is precisely a summation of many different aspects that makes a word to be considered to have a high degree of informativity (i.e. an “easy” word): e.g. that lexical and phonological contrast is high, predictability is high, articulation effort is low, and that perceptual contrast is high. In addition, these attributes appear to apply primarily to word-initial segments. Therefore, it seems essential to put the spotlight on the word-initial segments. What are the components of an effective articulation that at the same time are able to create maximum contrast for the listener to make use of? The primary goal is of course to be understood by the listener, and to enable contrasts, which will serve as lexical units. The speaker does this by utilizing the mobility in the oral cavity. However, we do not appear to utilize the full moving capacity of the mouth (Lindblom, 1983). Thus, the speaker moves her mouth less than is possible, as it is not the full capacity of the movements that determines the phonological contrasts. Furthermore, language is a system, and is likely a similar system to the speaker as to the listener, for the sake of efficiency. Thus, the speaker would likely utilize a system of articulation that interacts with the systemized structure of a particular language. In the word-initial segments, it is most important to get those structures and those movements right, to systematise them, in order to avoid misunderstandings and communicative collapse. One research question that has thus guided this dissertation is:

- *What systematic articulatory movements can we find in word-initial segments?*

To move the issue forward we need a more detailed theoretical framework.

1.2.2 Questions on the dynamic nature of speech

Research on systematic articulatory movements, and on the link between acoustics and articulation, is naturally focused on notions of what the phonological units are. Thus, we are looking for dynamic movement patterns that fit with predetermined structural units. But what if those units we traditionally believe to be phonological are not? Phonological structures that instead may be found in the body's own movements.

1.2.2.1 Towards a dynamical systems theory for speech

Dynamical systems theory (DST) is a collective name for mathematical differential functions of time and space[1]. Dynamical systems are thus measurable functions that are bound by relationships of different natural phenomena (for an overview, see Iskarous, 2016). Hence, they are physical laws, or laws of nature. One example of such a physical law is the falling object, where a stone released from a cliff falls faster and faster as a function of time. Another is the relationship between increase and decrease in an hourglass, again over time. We all interact with the physical laws, daily, all the time. Our speech, in turn, is based on the laws of nature. This is an indisputable fact: we make sounds with different movements governed by natural laws, and acoustics not only reflects these movements, it is also governed by those same laws.

Let us now return to one of the major challenges for linguists: to link the dynamic speech to a distinctive structure that can function phonologically and lexically. The critical point in understanding why DST is applicable to language is that in differential equations there is already a distinctiveness (Iskarous, 2017). Hence, we do not need to talk about how to bridge the gap (or explain the interface) between phonetics and phonology, between what is dynamic and what is discrete, because the gap no longer exists: linguistic structures are discrete and dynamic at the same time (Iskarous, 2017).

[1] The account of dynamical systems in this section stems mainly from a workshop with Dr Khalil Iskarous in Potsdam in autumn 2018, organized by Dr Aude Noiray.

The challenge therefore lies instead in mapping which dynamical systems are in the making, that is, what constitutes phonological structures.

One such system that is often applied in linguistic research is the damped mass spring systems (Saltzman & Munhall, 1989; Löfqvist & Gracco, 1999; Gao, 2008; Iskarous, 2016). Damped mass spring systems are based on the relationship between position, velocity (first derivative) and acceleration (second derivative). Quite simply, as a function of time, velocity is the change in position, while acceleration is the change in velocity.

In the simplified version (a linear oscillatory system), acceleration and position are the opposite poles; when acceleration is maximized, position is minimized (below zero) (see Figure 1). In addition, when velocity is at its highest, acceleration and position is zero (by "zero" is meant that there is no acceleration and that the position is in its initial position).

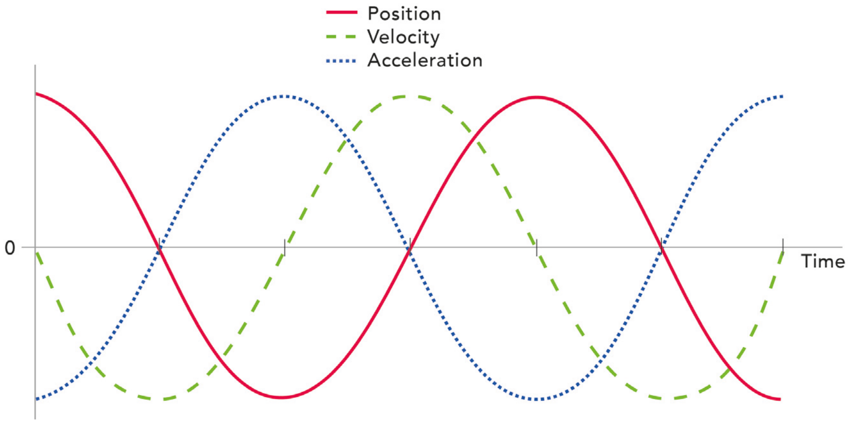


Figure 1. Linear oscillatory system
Sketch of the relationship between position, velocity and acceleration. When acceleration is maximized, position is minimized (below zero). When velocity is maximized, acceleration and position is zero.

A damped system (the damped mass spring system) involves a target; when a target is overshot, a force makes the movement move backwards. Thus, it involves deceleration of a movement as well, as a result of the target being overshot (see Figure 2). In other words, over time, a damped mass-spring system, without added positive force, returns to a stable equilibrium position. Mass-spring systems function much as car springs do, although car springs are not as flexible and changeable as human body parts (Hall, 2010).

Furthermore, the acceleration and deceleration of a movement are results of forces. Because of the overshooting of targets, internal forces decelerate the velocity. At the same time, external forces may not only maintain the velocity over time (if force ceases, velocity decreases), but perhaps also, as speed limiters, control how much the target is overshot.

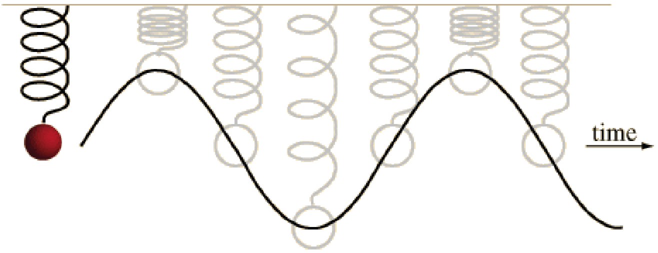


Figure 2. Damped mass spring system
An object's position moves as a function of time. If the object is attached to a string, and the string is stretched, the object begins to move. The object accelerates the most as it passes the starting point (equilibrium position), and then decelerates. Over time, without added positive force, the object will return to a stable equilibrium position. Image retrieved from: http://labman.phys.utk.edu/phys221core/modules/m11/harmonic_motion.html

Although the sum of the relationship between position, velocity and acceleration in a linear oscillatory system is always the same, its applicability to how the articulators move is complicated by the fact that several systems are presumably simultaneously active in speech. Internal force can, for example, be affected by the viscoelastic tissue law, and other possible non-neural factors, such as density or intra-oral pressure. Moreover, the oral cavity comprises several organs that may have different conditions: for example, the palate is robust while the tongue is extremely flexible with many degrees of freedom. Thus, the effect of an increase in speed will naturally not be the same for the jaw as, for example, for the tongue tip, not to mention how speed is adapted to, or controls, the constriction to be performed.

This complicated relationship may be further explained by dividing the dynamical systems applicable to speech into two categories: systems with *mechanical* properties and systems with *dynamic* properties (Perrier, 2012). The mechanical properties are, for example, the velocity, trajectory, and acceleration of an articulator, which can be recorded with available tracking devices. The dynamic properties, on the other hand, are the mechanical phenomena underlying the movement, such as external force, friction, or damping. These, which in turn can be divided into those that are more or less controllable via the Central Nervous System (CNS), and those that are intrinsic, are not as easy to measure as the mechanical movements (Perrier, 2012). Thus, although we are today able to measure the movements of the articulators, in order to fully understand them we need to put these patterns of movement into a context, or a model that includes the dynamic properties (Perrier, 2012). Hence, any phonological model of articulation should include and clarify, for example, the conditions of the oral cavity.

Applying the dynamical systems for speech is thus a complicated matter, as there are many systems, both mechanical and dynamic, operating at the same time in the oral cavity. This is further complicated by the task of finding the distinctive structures. Because, although dynamical systems, through the differential equations in mathematics, are distinct in themselves, it is a completely different matter to know what can function as a phonological unit. For this purpose, we need models and phonologies that clarify hypothetical structures and entities.

1.2.2.2 Where are the phonological structures?

Research shows that the same structure does not necessarily underlie speech motor control and, for example, limb movements (Perrier, 2012). This partly contradicts this dissertation's initial metaphor for articulation and acoustics: their similarity to the footstep and the footprint, respectively. Footsteps and articulation obviously exist under different conditions, the main one possibly being that the oral cavity is reasonably closed; therefore, adaptation is rarely made, during speech, to different environments (while the foot constantly adapts to changing circumstances). In addition, coordination between the parts of the speech apparatus is complex, and the relatively small body parts of the oral cavity make rapid changes during the course of the speech. As a result of extremely fast movements, feedback signals are less likely to work, which indicates that speech motor control may have local internal models, using the CNS (Perrier, 2012).

In a local internal model, it is understood that movements are based on different tasks to be performed. One model for speech that has this particular starting point is the *Task Dynamics model* (TD model) developed by Saltzman and Munhall (1989). The major features of the model are based on several years of motor control research (see further Saltzman & Munhall, 1989) that suggest that movements are truly coordinative structures guided by context-independent task-specific goals. The TD model, which is based on the notion of dynamical systems, further assumes that *gestures* are phonological units. Gestures, as described by Saltzman and Munhall (1989), consist of articulatory movements, which, in turn, are controlled by "speech-relevant goals", or tasks. Furthermore, each gesture unit is a synergy of muscles and joints to reach the gestural goals (Saltzman & Munhall, 1989). The TD model approach thus assumes that the articulators are relatively independent of each other, although they obviously cooperate in reaching the gesture goals.

The TD model predicts that gestures are selected at one level (*inter-gestural level*), and organized (for example, in time) at another (*inter-articulatory level*) (Saltzman & Munhall, 1989). Thus, without going into detail, this two-level model enables a feedback channel. Furthermore, when the tasks are performed within the inter-articulatory level, a contextual variation of the articulatory movements is still allowed. This is an important division in the TD model because it denotes how a given gesture (e.g. a bilabial gesture) is performed by a task of the lips (lip aperture, LA) with the help of various articulators (lower lip, upper lip, and jaw) (Saltzman & Munhall, 1989). The idea is that the task is to be performed independently of context (by so called *tract variables*, e.g. LA), while the articulators have some room to act. This is one reason why we may witness varied kinematic trajectories among speakers. It may also explain compensatory articulation, that is, articulation that is automatically reorganized (Saltzman & Munhall, 1989). It is further related to the idea of *via-points*, as proposed by Kawato et al. (1990), which allows dynamic variation while maintaining specific points in the movement (for an overview, see e.g. Perrier, 2012).

Because phonological units are gestures, which are thus assigned to different articulators with different tasks to perform, these gestures can overlap. Furthermore, gestures may display a temporal or/and a spatial overlap. Spatial overlap occurs when gestures share articulators and tract variables, while temporal overlap occurs when gestures are made with different articulators (Saltzman & Munhall, 1989). Regardless of the type of overlap, overlapping gestures serve as explanations for the co-articulatory patterns we see in speech, but are also able to explain many other phonological phenomena such as allophones, feature spread, speech errors, and speech disorders (Bell-Berti & Harris, 1979; Browman & Goldstein, 1989; Kent, 1997; Moen, 2006; Tilsen, 2016).

Furthermore, the TD model also assumes an external timing function, controlled through a *gestural score* (see further section 1.2.2.3 below). However, Saltzman and Munhall (1989) also maintain that even though the TD model assumes an external clock, the time function may possibly be another, which is more likely to be based on position and speed within the system's variables.

One final aspect of the TD model that should be mentioned here is that it assumes that the gestures are interconnected, and that the relationship between articulators must be specified in the coordinative processes of speech production (Saltzman & Munhall, 1989). This is considered to be through a coupling mechanism, which is defined within the various phonological units. Of course, how gestures are coordinated in time is based on their phonological representation, which will be addressed next in a gestural phonology.

1.2.2.3 A gestural phonology: Articulatory Phonology

The search for phonological structures, while guided by the dynamical systems, may have caused the emergence of gestural phonologies. One such phonology, developed in the late 20th century, is Articulatory Phonology (AP) (Browman & Goldstein, 1986, 1992; Goldstein & Fowler, 2003). AP is closely linked to the TD model and assumes that the phonological units are the abstract so-called *articulatory gestures* (not to be confused with co-speech gestures, e.g. manual gestures). These articulatory gestures follow the different task variables, as proposed by the TD model, tasks which they perform. Furthermore, the articulatory gestures in AP have specified start points and timing of targets (Browman & Goldstein, 1989, 1992). Therefore, it is important to note that time is included as a function in phonological representation. Other phonological theories often do not include a timing structure, which makes AP rather unique among phonologies (Kent, 1997; Turk & Shattuck-Hufnagel, 2020). Furthermore, the abstract articulatory gestures are hypothesized to be the fundamental units of both speech production and speech perception (Browman & Goldstein, 1986; Goldstein & Fowler, 2003). This is specifically inspired by the work of Carol Fowler on articulatory representation in perception and her theory of *Direct perception* (Browman & Goldstein, 1989; Fowler, 1986, 1996).

Using the framework of AP, the differences between two phonemes can be specified as follows: two bilabial sounds, for example /b/ and /p/, differ in that /b/ consists of one articulatory gesture while /p/ of two such gestures; in other words, the latter also includes a glottis gesture (Browman & Goldstein, 1986). As assumed in the TD model, the two articulatory gestures in /p/ overlap. Moreover, they are coupled with each other, and with gestures associated with adjacent sounds, in a spatio-temporal coordinated matter. Because of the coordination of multiple gestures to make a sound, there is no direct correspondence between a segment and a gesture (Browman & Goldstein, 1986). Of course, in AP there are many specifications for all possible phonemes, which in turn are language-specific, but the principle is the same: the phonological representation of a sound consists of several overlapping gestures linked to each other. How they are linked is in turn controlled by a *gestural score*, as mentioned.

A gestural score basically acts as a sheet of music, i.e. it specifies when in time an articulatory gesture should be performed (Browman & Goldstein, 1992; Saltzman & Munhall, 1989). Thus, it specifies the temporal overlap that occurs between the different gestures, as well as their individual duration (for visual examples, see Browman and Goldstein, 1989, 1992). A gestural score can in a simple way, for example, exemplify how a CV sequence has more gestural overlap than a VC sequence. In AP, the concept is adopted with an external clock that controls the time function of the coordination of the articulatory gestures. A gestural score can thus be said to effect this timing coordination. Although the mapping to acoustic segments is not one-to-one, the timing of the articulatory gestures can in turn manifest itself in acoustic patterns, especially in the form of different segment duration phenomena. One that can be explained in particular by the overlapping gestures is the *c-center effect*.

Byrd (1995) and Browman and Goldstein (1988) showe that the c-center effect is a phenomenon that arise due to the global organization of gestures. According to them, the c-center signifies the temporal center of the consonant's constriction (Byrd, 1995; Browman & Goldstein, 1988). This temporal midpoint is aligned with the onset of the following vowel. With only one consonant in the onset, you might say that the c-center is roughly the same as the acoustic midpoint (depending of course on the type of consonant). In clusters, each consonant's c-centers compete with each other, which leads to the mean value of all c-centers becoming the new c-center - thus the c-center effect arises (Browman & Goldstein, 1988). The coda consonant is instead "left-edged" with the acoustic vowel offset (Byrd, 1995). The c-center effect seems to explain why the vowel is different in length depending on the number of consonants in the onset, but not depending on the number in the coda position, although the pattern changes slightly when the cluster consists of more than three consonants (Byrd, 1995). These articulatory midpoints and c-center effects have been demonstrated in more languages than English (Kühnert et al., 2006; Marin, 2013; Marin & Pouplier, 2014).

The phonological interpretation of both the c-center phenomenon and the left-edged coda is that there are different connections between the various articulatory gestures. On the one hand, connections differ between onset (CV) and coda (VC), and, on the other hand, between consonants in themselves and vowels in themselves. This is where the hypothesis of "competitive coupling" comes in (Nam et al., 2009), that is, the observed timing patterns of gestures arise as a result of the different gestural onsets being either connected in-phase (simultaneous) or anti-phase with each other (a thorough description on the *coupling hypothesis* can be found in e.g. Nam et al., 2009; see also Gao, 2008; Mücke et al., 2012). However, it is still unclear how the consonantal and the vocalic gestures are connected. In the competitive coupling hypothesis the gestural onsets are connected. While the c-center effect is rather based on relationships between timing of consonantal targets to the vowel acoustic boundaries, or perhaps the vowel movement plateau, presumably also a target (the literature is ambiguous in this regard). In order for these two, onset-onset and presumed target-target relations, to be considered to belong to the same basic timing structure, a similar gestural duration of two consonants in a cluster is assumed, or the articulatory gestures that make up the consonants and the vowel may be doubly connected. In which case, uncertainties about how these two are related to each other clarify the need for more analyses of the CV relationship.

Some concepts in this thesis have been borrowed directly from AP. In addition to the articulatory gestures already mentioned, *tone gestures* are also referred to. Tone gestures can be said to be one type of articulatory gesture (consonantal and vocalic being other specified gestures) (Gao, 2008). From AP's point of view, a tone can be understood as articulatory gestural movements aimed at achieving a tonal task goal (Gao, 2008). They can thus be targeted as either a high tone, H gesture, or a low tone, L gesture (Gao, 2008). A tone gesture is thus a phonological unit that denotes the onset and target of a tonal movement; in other words, it is in principle an abstract signification of the vocal folds. Therefore, tone gestures are, in theory, as a unit, therefore interconnected with other articulatory gestures. Special interest has previously been taken in how tone gestures are linked to other gestures in different languages (Gao, 2008; Mücke et al., 2012).

In sum, AP presents a phonological model of the language that is based on the Dynamical systems theory, with use of the TD model. More details of the AP's framework are not given here, as it is neither the purpose nor within the scope of this thesis to specify gestural scores in the given language, Swedish. However, a further description of how tones have been shown to interact with consonants and vowels is provided in Paper 2. Further discussions are also given in the next section on Swedish phonology.

1.2.2.4 Research questions concerning speech dynamics

Before we get to the language under scrutiny, I would like to summarize some assumptions about dynamic speech, on which this thesis is based, as well as raise some related research questions. It is assumed that articulatory gestures, which are made up of articulatory movements, are phonological units, and that articulatory gestures overlap in time and in space. It is also assumed that articulatory gestures carry out specific goals, or tasks, or targets, as they are sometimes called. Another important assumption, related to this, is that articulatory gestures are timed with each other, for example at the start of the movements taking place. Furthermore, the articulatory movements rest on differential functions (e.g. damped mass spring systems), whose surface has only been scratched in linguistic research. The research questions that have guided the present dissertation are not intended to specifically test either the AP or the TD model but have been based on a genuine interest in how the articulatory movements are performed in time. The research questions on this subject can be summarized as follows:

- *How do you mechanically measure the time function of an articulatory gesture? What is its onset?*
- *Since acceleration is a result of added force - what phonological role might acceleration play in articulatory movements?*
- *How is the proposed interconnection of articulatory movements (e.g. onset-onset coordination in a CV sequence) affected by different intra-syllabic constraints?*

The last research question is further specified in the next section on Swedish phonology.

1.2.3 Questions on Swedish phonology

The following sections provide some background to Swedish phonology, with a focus on the Swedish word accent in section 1.2.3.1. One should keep in mind when reading the various studies that Swedish has a complementary vowel-consonant system (V:C and VC:, respectively), which is part of the Swedish syllabification rules. There is, however, a set of possibilities for the speaker to group the vowels and consonants in polysyllabic words. We have, for example, 1) long vowel followed by an internal juncture, 2) long vowel followed by a mora-sharing short consonant (internal juncture ends up in the middle of the consonant), 3) short vowel followed by geminate (the internal juncture divides the geminate), 4) short vowel followed by short consonant (sometimes occuring as mora-sharing), 5) short vowel followed by short consonant and then by a mora-sharing geminate (for a more detailed description, see e.g. Gårding, 1967, or Riad, 2014). Thus, consonant clusters can occur both word-initially (up to 3-member sequences) and post-vocalically (up to 4-member sequences) (Sigurd, 1965). However, consonant clusters are rare in word-medial position after long vowels. Moreover, long consonants, geminates, occur only in a word-medial position, and are considered to occur only together with short vowels. However, example 2 above suggests that the mora-sharing coda may be part of a prolonged consonant.

The Swedish phoneme system is considered to consist of nine vowel pairs (long/short) and 18 consonants, most of which occur as long and short variants (Bruce & Engstrand, 2006; Riad, 2014). For this dissertation, the short vowels [a] and [ɪ], and the long vowels [ɑ:] and [i:] are of particular interest. As indicated, a difference in vowel quantity also means a difference in vowel quality in Swedish. For the high vowel /i/ the change in quality is not as great as for the low vowel /a/, where the tongue moves significantly backwards when quantity increases (Bruce & Engstrand, 2006). Further information on the consonants and vowels selected for study can be found in Chapter 2, Methods.

Furthermore, Swedish has several stress levels. In addition to stressed/unstressed, main stress is the highest level. In compounds, a lower level of stress can be placed on a subsequent syllable (Elert, 1964). But in general, one stress per word applies; its placing is specified by morphology (Riad, 2014).

1.2.3.1 Swedish word accent

Words in Swedish have one of two tonal accents, known as the Swedish word accents (SWA): Accent 1 (A1) or Accent 2 (A2). They are sometimes also referred to as *acute* and *grave*, respectively. They increase the prominence level of the stressed syllable and are often considered to include several tones (Elert, 1964; Bruce, 2007).

SWA are both morphologically and phonologically different to each other (Öhman, 1967). Both A1 and A2 display a tonal peak, but the timing of the peaks differs so that A1 has an early tone peak in the word compared to A2. This timing difference of the high tone is found in most Swedish dialects, although it can be realized as different tones in word-initial position. These word-initial tones, or stem tones, are used in prediction of upcoming words (Roll et al., 2013). The word accent is said to be induced by the suffix (Riad, 2014). Thus, the definite form /ˈbiːlɛn/ (the car) gets an A1 stem tone, while the indefinite plural form of cars /ˈbiːlar/ instead gets an A2 stem tone, and so does a compound: /biltvätt/ (car wash), /bilstol/ (car chair), and so on. Hence, A2 has more possible continuations than A1. In South Swedish, which is the dialect variety investigated in this thesis, a word-initial high tone is a cue for A1, while a word-initial low tone a cue for A2 (Gårding & Lindblad, 1973; Bruce, 1977; Roll et al., 2013).

The whole combined tonal pattern of A1 is often considered to be early in the word while in A2 it is late in the word. The so-called pulse model (Öhman, 1967) describes these timing differences as a result of a negative pulse constituting the word accent, which thus creates the physiological property of the tonal fall (Öhman also makes a link to a glottal stop, which occurs in the Danish *stød*). According to the pulse model, this negative word accent pulse occurs simultaneously as a positive pulse, which is responsible for the overall tonal rise. The positive pulse is of a higher prominence, which Öhman (1967) suggests is a kind of basic phrase contour. The negative pulse is thus superimposed on the basic phrase contour and breaks off the tonal rise that is the presumed result of prominence. Öhman's intonation model (1967) has a clear connection to the physiological properties of f_0, that is, that the negative pulse causes a break in the tension in the vocal folds. Thus, it is a predetermined order in time of a longer positive pulse with a simultaneously shorter negative pulse that is assumed to create the rather melodic tonal pattern in Swedish (as perhaps particularly evident in the pattern of two tonal peaks in a row in high prominent A2 in some Swedish dialects).

According to Öhman (1967), the timing differences between the SWAs are due to an early negative pulse in the word in A1 but a late negative pulse in A2.[2] Unfortunately, all word examples in Öhman's study are based on so-called "sentence accents". Thus, in his intonation model, the different prominence levels of the word and sentence accents are not separated. Bruce (1977) aptly showed that the tonal patterns of high-prominent sentence accents were systematically different than low-prominent word accents. Without going into too much detail, it challenges the pulse model, and the idea that word accents are superimposed on basic phrase contours. Furthermore, it is not clear whether the SWA can be said to consist of only a negative pulse (although Bruce in his model, 1977, also signifies the SWA as tonal falls with different timing). However, Öhman's pulse model may be applicable in a slightly transformed form, which will be discussed shortly.

1.2.3.2 South Swedish

The tonal timing difference in South Swedish word accents is visualized in Figure 3. As can perhaps be read from the figures, the auto-segmental representation of the low-prominent word accents in South Swedish is for A1 based on a tonal fall in the stressed syllable (H*L), and for the A2 a tonal rise (Bruce, 2007), or an LHL pattern (Riad,

[2] However, according to Öhman (1967), the order of the Malmö dialect is reverse to that of central Swedish, meaning that A1 instead starts later than A2. The negative intonation pulse begins only after the consonant, which creates a tonal fall during the vowel. In A2 the negative intonation pulse instead occurs already while the first consonant occurs, that is, it gives a tonal increase (i.e. the positive intonation pulse) during the vowel.

2006) within the stressed syllable (L*HL). Furthermore, the SWA are usually considered to be linked to the syllable (Bruce, 2007). As is evident in Figure 3, the tonal fall in A1 stabilizes at the syllable boundary. The tonal fall in A2, on the other hand, does not seem to stabilize until in the second syllable, although most f_0 changes takes place in the stressed syllable in both accents.[3]

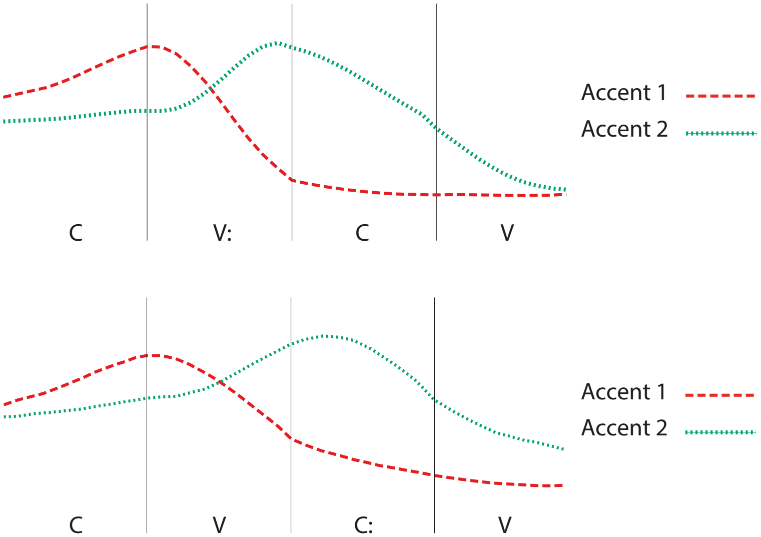


Figure 3. Swedish word accents
Visualization of the word accents based on mean f_0 of 19 South Swedish speakers (male and female speakers combined). Open stressed syllables (CV:.CV) with long vowels (top): /ˈmɑːnɛn/ (A1) - /ˈmɑːnar/ (A2); and closed stressed syllables (CVC.CV) with short vowels (bottom): /ˈmanːɛn/ (A1) and /ˈmanːa/ (A2). Segment duration is normalized.

That the tonal activity is not completely limited to the first syllable in A2 may be due to its purpose. Of the two SWAs, A2 is considered to be lexical, and is often called a *connective accent* because of the tone-inducing morphological rules, which enables more continuations (A1 is referred to as an *isolated accent,* as monosyllables often carry A1) (Bruce, 2007; Riad, 2006). Furthermore, incentive from a perceptual study on A2 indicates a tri-tonal accent (Ambrazaitis & Bruce, 2006). Indeed, the tones of SWA have been proposed to bear different roles, which in the South Swedish dialect would

[3] This might depend on prominence level. As Gårding and Lindblad suggest (1973), the southern Swedish tonal pattern is more adapted to morpheme boundaries than to syllable boundaries. This is possibly due to the prominence level, as Gårding and Lindblad (1973) in their study only discuss word accents with different focus types. Indeed, in a pilot study, we noticed that the f_0 activity in the high prominent /ˈbiːlɛn/ and /ˈbiːlar/ followed the morpheme boundaries (based on a reinterpretation of the data, as it initially contained errors) (Svensson Lundmark et al., 2015). However, for low-prominent words, the syllable indeed seems to be bearing.

mean that the last L tone is a boundary tone, that the H tone before signals prominence, and that only A2 has a lexical tone, which would be the first L tone (Riad, 2006). Thus, the two SWA might not only differ morphologically and phonologically, but also lexically from each other.

1.2.3.3 Indications of tone gestures in Swedish

If tones are to be converted to tone gestures (as proposed in Gao, 2008), then A1 in South Swedish would presumably consist of two gestures; an H gesture (a rise towards a high tone) followed by an L gesture (a fall towards a low tone). A2 would instead presumably consist of three gestures, in sequence, an L gesture, an H gesture, and another L gesture. Although, as shown in Figure 3, the initial L gesture in A2 (the proposed lexical tone) does not appear to include a fall, like the other two low tone gestures in SWA do (which are incidentally suggested to be boundary tones). The initial L gesture in A2 would instead presumably be a result of overlapping tone gestures (similar to the proposal for Mandarin tone 2 by Gao, 2008).

However, the idea of tone gestures is hypothetical. For example, it is unclear whether a tone gesture is characterized by its onset or its target, or by both. It is also unclear whether f_0 is a reasonable representation of a gesture. Neither is it certain how the onsets and the targets of a tonal fall and a tonal rise, respectively, are controlled by the speaker.

As is well known, f_0 control is a combination of sub-glottal pressure and laryngeal muscles that control tension in vocal folds, and the cricothyroid (CT) muscle is generally considered to be primarily responsible for f_0 elevation (see e.g. Hirose, 2013; Erickson, 2013). CT activation and f_0 do not follow each other exactly and there is a time delay for the resulting f_0 signal (Hirose, 2013). An EMG study on SWA in particular measured a delay of about 80 ms between EMG and tonal peaks (Gårding et al., 1975). If one wants to examine articulatory coordination, in a comparison of tone gestures, acoustic f_0 may therefore not be appropriate.

What is also problematic about making f_0 substitutes for an articulator (as in tone gestures) is that f_0 lowering seems to involve other physiological functions than f_0 rising. Although it is not clear what the actual functional onset is for low f_0, a lowering of the entire larynx is involved, which includes the extrinsic muscles and attached structures such as the hyoid bone (Erickson et al., 1983; Honda et al., 1999; but for an overview see e.g. Hirose, 2013). In either case different articulatory activities are included in an f_0 rise compared to an f_0 fall. We should therefore assume that these are different gestural activities. Thus, a phonological articulatory model that includes tone gestures should have a distinction between not only gestural onset and targets, but also a distinction between tonal rise and tonal fall, respectively (also alluded to by Gao, 2008).

To return briefly to Öhman's pulse model. Öhman (1967) has suggested that the SWA consist of a consonant-like gesture, similar to glottal stop. This is reminiscent of the idea of tone gestures from the AP framework, since Gao (2008) suggests that tone gestures in Mandarin act similarly to consonantal gestures in their coordinated relationship to the vowel. Thus, a possible interpretation of SWA is that they too function like consonantal gestures in their relation to the vowel. More similarities between the notion of tone gestures and Öhman's pulse model (1967) can be found in the fact that Öhman divides f_0 rise and f_0 fall as depending on separate pulses, which are activated in time and can be in succession or overlapping. Thus, similarities are also found to the overlapping articulatory gestures in the TD model (Saltzman & Munhall, 1989), and the proposal in AP on tone gestures with different targets for high or low tone (Gao, 2008). Thus, there are several incentives for coordination between articulators that create f_0 and those that create consonants and vowels.

That SWA would consist of tone gestures is a new idea that has not been tried before. However, there are some question marks. As it is not completely established how the tone gestures are realized, the investigative part of the dissertation on the Swedish tones is limited. This dissertation thus does not look at the individual tones of the SWA, but rather at how their presence affects the surrounding and simultaneous articulation. If we are really dealing with tone gestures, we should see evidence of their presence in the other articulatory movements. Adopting this approach, we might also be able to clarify whether f_0 is a sufficient representation of the tone gesture. In any case, it is hoped that the results can contribute to a nuanced discussion about alignment, hierarchical affiliation and how tones are coordinated with articulation. Research questions that have guided the studies on this topic can be summarized as follows:

- *How do tone gestures manifest themselves in Swedish?*
- *How do falling and rising tones affect the articulatory movements at word onset?*
- *How is the connection between the consonant and the vowel affected by different tonal contexts?*

1.3 Limitations of previous studies

This dissertation presents a variety of articulatory measurements. These are presented partly in the various studies, but also recur in the discussion (Chapter 4) where, among other things, a review of some methods is made. Many previous studies are based on different articulatory measurement methods, which makes these results difficult to compare. For example, there is no consensus on how to measure onset-onset time lags, which is necessary to make cross-linguistic comparisons. Sometimes a basic understanding of mechanical movements seems almost lacking, which can make a phonological analysis unclear if not completely incorrect. An important purpose of the dissertation has thus been to understand how some of these basic measurement methods are connected.

Another limitation of previous research may be the number of speakers. Since we still know relatively little about the dynamic structure of speech, and the variation seems to be great, it is tempting to do case studies of a few speakers. However, these analyses can lead to hasty conclusions, as the variation is not only great between speakers, but also within speakers. Until we know what is consistently and systematically varied across speakers we might need to include as many speakers as possible. An intermediate goal of the dissertation has been to collect data from a large number of speakers, since it is only through such an analysis that we can find the systematics in the dynamics.

The Swedish word accent has attracted scholars for a long time and from many perspectives: production-based (e.g. Öhman, 1967; Gårding et al., 1975; Elert, 1964; Löfqvist, 1975); perceptually (e.g. Ambrazaitis & Bruce, 2006; Roll et al., 2013); phonologically (e.g. Bruce, 1977, 2007; Riad, 2006). The dialect typology in particular has gained interest since the Swedish dialects show varied tonal timing patterns between both the dialects and between the accents themselves (Gårding & Lindblad, 1973; Öhman, 1967; Bruce, 2007). Studying the dialects can provide insight into how SWAs in particular are modelled, but also shed light on tones in language in general. Although most phonetic orientations mentioned have reasoned about articulatory differences between the two SWAs, there are few articulatory studies today. Thus, a large part of the thesis is dedicated to examining the Swedish word accents from an articulatory perspective which has not been done extensively previously.

1.4 Scope of the dissertation

As the dissertation title suggests, this thesis explores some word-initial segments. These segments are limited to the CV sequence, which may be considered as somewhat of a universal model of the syllable. Since three of four studies examine the word accents, which are characterized primarily by f_0, it was initially considered important to work with an unbroken f_0 curve. Therefore, the choice was made to examine the voiced CV sequences /bi/ and /ma/ (thus, the consonantal gestures consist mainly of bilabial closing and opening gestures, while the vocalic gestures are both palatal narrow and palatal wide). This might be considered a limitation regarding vowel and consonant combinations, but the focus in the dissertation on these particular CV sequences is still appropriate. Many other studies have examined the CV sequence /ma/ in particular, which makes it useful for comparison. Second, studying gestures made with two different organs allows us to focus primarily on temporal overlap and avoid, as far as possible, spatial overlap that further complicates matters. Third, one of the first contrasts a child learns has long been considered to be that between a nasal stop and an oral stop: /ma/ vs /pa/ (Jakobson & Halle, 1956). Thus, a word-initial CV sequence consisting of a bilabial consonant and an open vowel is one of the most basic building blocks of language and therefore highly suitable for speech production model studies.

The studies are mainly based on recordings with articulography. Learning this methodology has been one of the sub-aims of the dissertation project. The acoustic analyses are also made on these combined kinematic and sound recordings. There is no doubt that with full acoustic recordings (outside the Articulograph) the results would have been different, since the speakers would not have had objects glued to their articulators. On the other hand, a spontaneous speech in a natural setting would most certainly have produced other results as well. Speaker situations are crucial, and in this dissertation, all recordings have been made in a lab environment, which should be taken into account when reading the work.

Initially, Papers 1 and 2 were based to some extent on the theoretical framework of which AP is a part. Concepts have been borrowed from both the TD model and AP. The studies deviate from several aspects of AP's framework and can therefore rather be said to be inspired by its phonology. Similarly, direct aspects of the TD model are not tested, but the results from all four studies may be interpreted based on the model. What runs like a main thread throughout the studies, however, is the approach to the dynamical systems. The dissertation is therefore to be considered as an exploration of the possible applicability of DST to speech production models.

1.5 Outline of the dissertation

The thesis consists of four papers and an introductory chapter. Three of the papers investigate the Swedish word accents, with each study a more and more specific approach. Paper 1 has the nature of a pilot study with a multi-dimensional acoustic and articulatory perspective. It functions as an introduction to the field. Paper 2 is explorative, and a thorough and specific examination into the articulation of the lips and the tongue during a word-initial CV syllable onset comparing the two SWA. The paper that follows complements Paper 2 with results on the jaw movements. Paper 3 presents a novel approach to articulatory measurements and thus stands on its own. The fourth and last paper returns to the acoustic and holistic perspective that could be seen in Paper 1, but it does not investigate the SWA. Instead Paper 4 looks at how consonant articulation affects non-adjacent consonant segments.

2 Methods

The following chapter discusses speech material, speakers, procedures, measurements and statistics as used in all four studies. All data has been recorded with electromagnetic articulography. To help the reader form an understanding of what it is like to work with articulography, those parts of the chapter that refer to decision making are described in detail. The section on procedure is specifically focused on presenting articulography, as well as describing how recordings are made. A large part of this chapter also aims to describe the various measurements used, including their technical implications and methodological benefits. A more thorough evaluation of some of the, in my eyes, more useful articulatory measurements can be found in chapter 4 (section 4.3).

2.1 Speech material

An early aim of the dissertation project was to investigate articulatory differences between the Swedish word accents: A1 and A2. To explore this, data collection began with a two-person pilot study. The large corpus data that followed included kinematic and acoustic data on 21 speakers.

2.1.1 The pilot study

Since monosyllabic words get A1, polysyllabic words must be included in any comparison between the word accents. For the pilot study, the decision was made to investigate simplex disyllabic target words, which can carry both word accents. The disyllabic A1 and A2 target words all had stress on the penult and were produced in the carrier phrase "Det var TARGET jag sa." (*It was TARGET I said.*) (Appendix I). Thus, in the sentence context the target words elicited a narrow focus accent. Moreover, the words were also matched so that each A1 – A2 pair consisted of the same nominal word stem, but with different suffixes: definite singular for A1 and indefinite plural for A2.

The target words varied in quantity (long/short) and in vowel quality: low /a/, high /i/, and rounded /o/. Less energy was spent on the choice of consonants that varied as a result of getting the vowels needed. However, voiced consonants were of interest in order to analyse an unbroken f_0. Each vowel quantity pair (V/V:), was matched with an ensuing consonant (C2). Thus, the post-vocalic /l/ was matched in the target words: /ˈbiːlɛn/-/ˈbiːlar/ and /ˈbɪldɛn/-/ˈbɪldɛr/, and in /ˈvɑːlɛn/-/ˈvɑːlar/ and /ˈvalːɛn/-/ˈvalːar/; and the post-vocalic /v/ was matched in /ˈvuvːɛn/-/ˈvuvːar/ and /ˈbuːvɛn/-/ˈbuːvar/.[4] Furthermore, the target words also conditioned either an open or a closed stressed syllable (CV:.CVC or CVC.CVC), which in turn contained either a long or a short vowel. Thus, the material consisted of twelve target words, all matched as word accents. The material conditioned accent type, syllable type, consonant type, vowel type and vowel length.

Two female speakers of South Swedish (age 38 and 49) read the material ten times each, i.e. 80 target words per speaker (see further details in section 2.2.3). They were both researchers in linguistics and conversant with the purpose of the study. During the recordings /ˈvuvːɛn/-/ˈvuvːar/ proved to not differ in terms of word accent (both were elicited as A2) and therefore those words along with their matching counterpart /ˈbuːvɛn/-/ˈbuːvar/ were not considered further in the analyses that followed. A subset of the data from the pilot study was used in Paper 1, together with data from a different project: Function- and production-based modelling of Swedish prosody (Swedish Research Council, 2009-1566, PI Johan Frid).[5]

2.1.2 The corpus

The corpus material was designed to examine articulatory differences between the SWAs. To make sure that the effect of the tonal context on articulation was an effect of word accent, the target words were placed in a low-prominence context, that is in target sentences that followed leading questions. The leading questions ensured that a narrow focus (contrastive) was placed on the last element of the target sentence instead of on the target word.

First drafts of the corpus included complex onsets and codas compared to singleton onsets and codas for the possibility of addressing the c-centre hypothesis (see section 1.2.2.3). This material was tested in the AG501 in December of 2016. However, the many different conditions, combined with the goal of putting the target word in a low prominence induced context, made for complicated sentences for the speaker to read. The kinematic data became hard to analyse with many exceptions and deleted sweeps as a result. Because of this, and because of the scope of the dissertation project, focus

[4] The transcriptions are phonematic, however, displaying major allophonic variation for vowels related to quantity.

[5] A subset of the pilot study data was also used in a conference paper (ICPhS 2015), but due to a fault in the analysis the results was misleading (Svensson Lundmark et al., 2015).

was on comparing only the effect by the two tones on simplex articulatory movements (instead of coordination of many articulators, which may be the case for clusters).

It also became evident that closer attention needed to be on the choice of articulators and articulatory movements. To measure the articulatory movements in a close vowel is complicated because of the small differences in movement in the narrow opening while, for a rounded vowel, the inter-gestural articulation of lip-rounding and tongue body movement also make for a difficult-to-handle analysis, because of the many variables. In addition, coronal consonants in combination with a close vowel create difficulties for large data handling, as manual adjustments become more necessary. For these reasons, bilabial consonants in combination with an open vowel were chosen to be the hub of the corpus material. Moreover, many previous studies used bilabial consonants in combination with open vowels, which made it useful for comparison. Other consonants and vowels were also included so as to make possible intra-articulatory coordination comparisons (Appendix II). Furthermore, only voiced consonants were included (even if f_0 would still be influenced by the consonant, e.g. the murmur in the nasals).

Another aspect that was tested in a pilot recording was the possibility of using non-words (e.g. Mámi/Màmi) in identical sentences, in order to have control over the segmental structures. However, it was problematic for the speakers to get the right word accent without a proper context or suffix as a guide. Being connected to an Articulograph in a lab is stressful enough. Participants need easy sentences to read, otherwise they might produce more saliva causing the sensors to fall off. Thus, in addition to the many obvious benefits of using real words, it was also important to avoid potential artificial effects.

Using real words, in turn, meant that the target sentences ultimately used were not identical. However, the target sentences had a similar information structure, and segmental structure was as similar as possible: the target word was preceded by 5 syllables and was followed by 3 syllables, the last two of which consisted of a word with a narrow focus. In addition, the three adjacent segments before the target words consisted of the identical suffix /adɛ/. An example of a target sentence that followed a leading question is: "Var lämnade Inger malen hon fångat? Inger lämnade malen till kocken." (*Where did Inger leave the catfish she caught? Inger left the catfish to the cook.*) (target word underlined).

The target words were as follows (in A1 – A2 pairs): /ˈmɑːnɛn/-/ˈmɑːnar/, /ˈmanːɛn/-/ˈmanːa/, /ˈmamːɵt/-/ˈmamːa/, /ˈnanːʏ/-/ˈnanːa/, /ˈnamnɛn/-/ˈnamnar/, /ˈamːɛn/-/ˈamːa/, /ˈbɑːlɛn/-/ˈbɑːlar/, and /ˈmɑːlɛn/-/ˈmɑːlar/. Thus, in addition to the SWA, the material conditions a variety of variables: vowel quality ([ɑ] – [a]), vowel quantity

(V: - V), consonant quantity (C: - C), consonant type, syllable type (CV: - CVC), etc.[6] Papers 2-4 are based on different subsets of the data as a result of these different possible variables. The full corpus speech material with target words underlined is found in Appendix II.

2.1.2.1 Speakers

Since tonal patterns and articulation differ between Swedish dialects, speakers included needed to speak the same dialect. For practical reasons, and also because South Swedish was understudied, speakers from this region were chosen. That is, South Swedish speakers currently residing in Scania, who, more importantly, considered themselves to be speakers of the South Swedish dialect "skånska" (≈ Scanian). 21 speakers participated in the recordings. The speakers' age ranged from 23 to 75 (the average being 40 years, and the standard deviation 12,3 years). All participants (but one: the author) were ignorant of the purpose of the study. The speakers were collected by advertisement in an internal newsletter, by word of mouth, and by posters at the Centre for Languages and Literature (see Appendix III). No payment or reward was offered for participation. The speakers come from different parts of Scania (see Figure 4) that differ somewhat in terms of dialect. The difference between the dialect regions that might be relevant for this thesis occurred in the long vowel [ɑ:]: dipthongization in the south and the southwest, and almost no dipthongization in the north.

16 speakers grew up with both parents speaking South Swedish, four of the speakers with one parent speaking another L1 than Swedish. In one case, neither parent spoke South Swedish, and this speaker was the only one who did not display the typical South Swedish tonal pattern of the SWA, with a tonal fall in the stressed vowel for A1, and a rise in the stressed vowel for A2 (instead both were A2 patterns). The speaker with the deviant tonal pattern was excluded from the studies. Another speaker read only the first draft of a speech material (before alternations to the final script); hence could not be included in the analysis. The speech data of these two speakers (with the deviant tonal pattern, and the deviant speech material) is still available for analysis in the corpus.

2.1.2.2 Observed speaker variation due to the material

Although prominence was controlled in the corpus material by using leading questions, some observed variations occurred both during the recording and during the initial

[6] In Swedish, stressed syllables are predominately either CV: or CVC:, meaning that a short vowel is generally followed by a long consonant (Riad, 2016). However, there are divided opinions as to whether there is a strong consonant contrast in South Swedish or not: although the vowel contrast is high in South Swedish, the C:/C ratio is low (Gårding, 1974; Garlén, 1988; Schaeffler, 2005; but Riad, 2016). It follows that, in this dissertation, geminate and long consonant are denoted C:, except when the long consonant is included in a cluster. There it is instead referred to as a short consonant, CVC.CVC (e.g. /ˈbɪldɛr/ instead of /ˈbɪlːdɛr/).

calculations of the data points. In connection to the speech material and the recordings this variation will now be discussed in more detail.

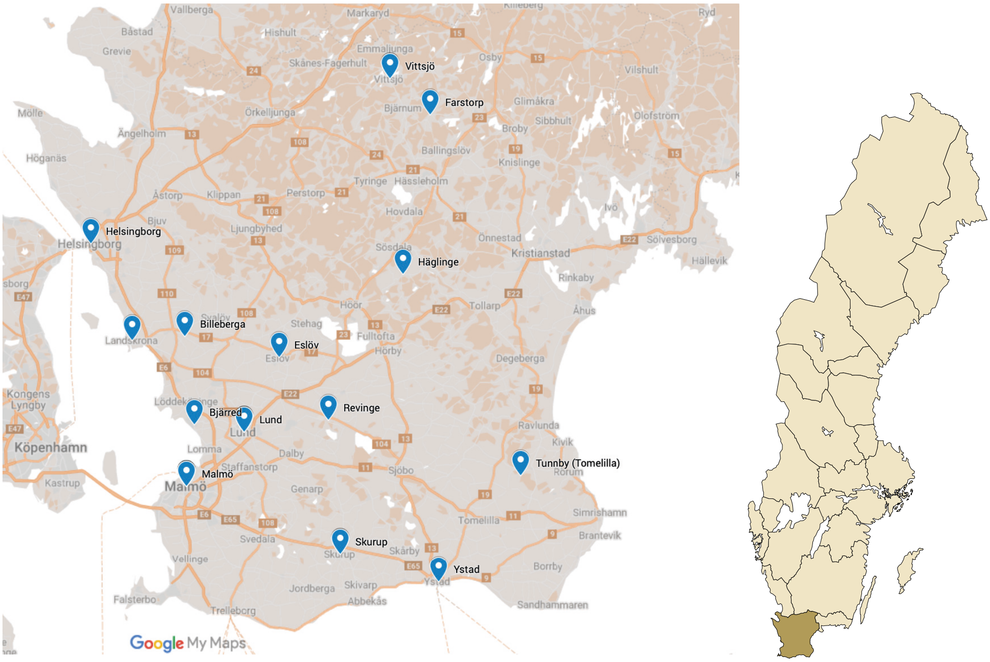


Figure 4. A map of Scania.
The map shows the approximate different locations of speaker's origin, see also Appendix IV (Google, n.d. Retrieved from https://www.google.com/maps/d/u/1/edit?hl=sv&mid=1YgPLJdsyKYes4BVP6f-NMqgYVefb0o13&ll=55.95214612170872%2C13.01303152703554&z=9).
Image of Sweden (right) (Retrieved from: https://commons.wikimedia.org/w/index.php?curid=7028772)

One observation that was made during the recording was that speakers wanted to mix things up, possibly due to tediousness. As a result, target words sometimes received higher or altered prominence. Although the target words were in random order for each speaker (each speaker had a unique set of the order), each "set" was repeated with the individual speaker, which may have had an effect on the recording. However, random prominence seemed to occur in all words and figure in most of the speakers. Therefore, this is presumably sufficiently dealt with in the statistical model.

However, the speech material might have been constructed involuntarily in such a way that it opened up for systematic deviations from the intended prominence patterns. There are a few words that stand out especially because their information structure is different from that of the other sentences. These are the target sentences that read in the following way: *Isak rhymed 'urge' with 'swans'* (in Swedish: “Isak rimmar ‘manar’ med ‘svanar’.”) (see Appendix II). The target sentence is an answer to the leading question: *What did Isak rhyme ‘urge’ with?* This target sentence seems sometimes to attract higher *f*₀ by some speakers. In some instances, the speakers also insert a short pause after /ˈmɑːnar/, thus signaling a phrasal boundary. Hence, the target word /ˈmɑːnar/ attracts often more prominence pattern corresponding to a phrasal accent than a sentence including a target word such as /ˈmɑːlar/ (*catfishes*), as in: *Where did Åsa leave all the catfish she caught? Åsa left catfishes by the boat.*

Besides /ˈmɑːnar/, /ˈamːa/, /ˈnanːʏ/ and /ˈnamnar/ also have a similar information structure, the target word following the verb *rhyme*. It could be that this particular sentence structure "X rhymes TARGET with CONTRASTIVE-FOCUS" enables the possibility of two choices for the naïve speaker. The speaker could either read it as one phrase, thus, place all prominence on the CONTRASTIVE-FOCUS (that is, a narrow focus), which was the intention of the researcher when constructing the speech material. The speaker could also decide to divide the sentence into two phrases and place phrase-final prominence on both the TARGET and the CONTRASTIVE-FOCUS. A third option for the speaker is to place a broad focus on both target and the CONTRASTIVE-FOCUS, which has been observed in the case of one or two speakers. However, this third case is not limited to the so-called *rhyme*-words but is equally distributed between the target sentences in the case of the speakers who had this strategy. In Paper 4, a comparison is made between some of these target words, in an attempt to determine whether prominence had an effect on the results.

2.2 Procedures

The procedure and the experimental setting were almost identical in the recordings made for the pilot study and for the corpus, which was done using articulography. Both recordings took place at the Lund University Humanities Lab, an inter-disciplinary department for research technology and training at the joint Faculties of Humanities and Theology at Lund University (www.humlab.lu.se). Before we go into procedural details, a presentation of some of some of the advantages and disadvantages of articulography seems called for.

2.2.1 Articulography

Articulography enables tracking by means of sensors glued to the speech articulators during speech. By means of carefully placed sensors (attached with cords) on selected articulators, we can measure how each part of the mouth interacts over time and space. In addition, it is possible to simultaneously record acoustic and kinematic data, as well as to obtain output in the form of 3D visualization and large amounts of data in several dimensions. As the tracking of the sensors is fast, articulography is particularly suited for timing analyses, especially timing that occurs between articulators.

Articulography has several names and it is sometimes also referred to as EMA, which is short for ElectroMagnetic Articulography. The data in this dissertation was recorded using an EMA by the Carstens company; an AG501 (Figure 5). In the Carstens AG501, there are nine transmitter coils inside the three arms, which hang over the head of the speaker, seated below. These transmitter coils emit magnetic fields at different frequencies. By means of an alternating current of sensors (glued to the articulators of the speaker), a distance to each transmitter coil can be obtained. It is then possible to obtain x, y and z coordinates, as well as two directional angles on each sensor. The sampling rate is 250 Hz/1250 Hz, and at the EMA in Lund there are 16 channels, i.e. the possibility to record data from 16 placed sensors.

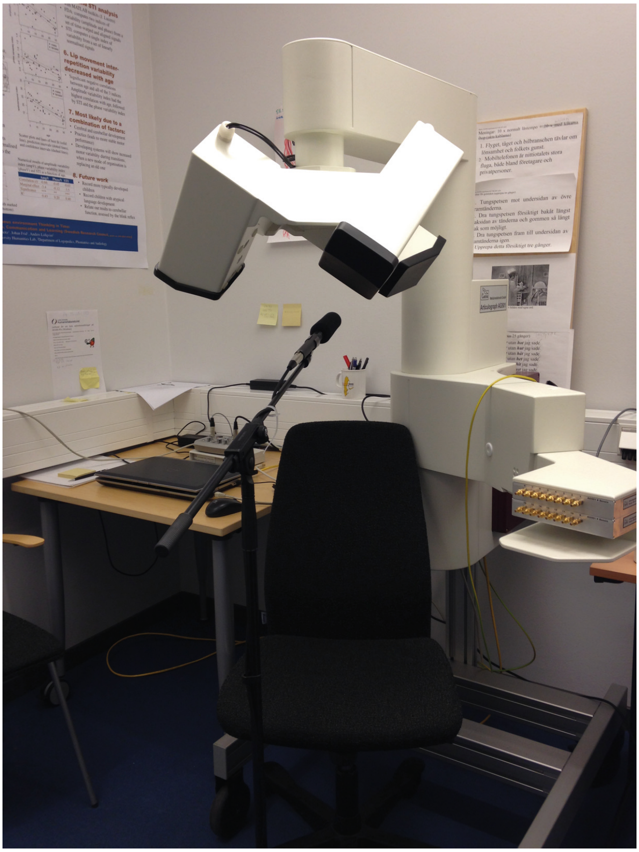


Figure 5. The AG501 located at the Lund University Humanities Lab.
The participant is seated in the chair below the three coils. The coils are connected to the device to the right (there is room for 16 sensors/positions which are connected simultaneously). The microphone is included in the photo. Dataprocessing is done on the laptop to the left, where the experimenter also overviews the recordings in real time. The prompter, on which the speech material is shown, is 2-3 metres behind the experimenter when seated (out of picture, to the left), in eye-level of the participant.

There are pros and cons with working with a device like an EMA. The main cost of the EMA is that it's invasive, that is, its presence may affect exactly what you want to explore. Thus, the cords that are attached to the sensors have a certain effect on speech and swallowing. Even though the speech apparatus of the participant most often gets used to the cords after a while, the surroundings, the experimental setting, may also affect the ease of the participant. In that sense, the lab setting itself can be invasive. Thus, like any lab setting it has a high internal validity.

Another cost of the EMA is that it is time-consuming. It requires a lot of preparation, and during the experiment the sensors might fall off, which is not only time-consuming but interrupts the concentration and the flow of the reading by the participant. Moreover, the experiment leaders and lab assistants need some fundamental skills, such as knowledge of the computer software or of the device itself, routines on equipment and with the participants etc. These fundamental skills can partly be taught, but are mainly learnt by experience, which obviously takes time.

Although the 3D output that one can get from the EMA is compelling, the observation itself is however not "the experiment" (as in the sense of a quantitative analysis). The experiment is centred on the data output, which is a matrix of movement data. Since the recordings are made in 250Hz this means collecting the position of each sensor every 4 ms, hence, 250 times each second. As each speaker may read for about 20 minutes (1 200 seconds), and each sensor is recorded in five dimensions, with eight sensors, this means a table of approximately 12 million entries – for one speaker. As is easily understood this huge amount of data is of less use if the measuring technique is not mastered. Measuring enables a higher degree of testability, but the decision on what measurement to use will in turn affect the degree of testability (Popper, 2002). Thus, understanding what to do with the data is essential, which is simply impossible without following some kind of speech production model, or looking at what others have done in the fields of phonetics and phonology.

One of the benefits of articulography, however, is that it is an "evoking device", that reduces the non-sensible to sensible (Hacking, 2010). With the EMA we can visualize both the things we cannot see, the unobservable entities inside the oral cavity, together with the things we can see but are too fast to observe, e.g. the visual orofacial movements. They are treated equally and the data on them can be processed at the same time in the same ways. Thus, together with the simultaneous sound recordings we can make measurements of many modalities at the same time.

What is most special about EMA is that it has a huge potential. We know that articulators are "real", and that they cause speech sounds. With an articulograph we can study in detail how the articulators move as we speak, and, with phonological knowledge, we can investigate, for example, the overlapping pattern of consonants and vowels. Thus, instead of measuring the acoustic imagery of vowels and consonants, much like footprints, we can look at the very cause of the imprint - the footstep - or in this case, the movements of the articulators. An articulograph gives us clear and rich data on a how a body moves: its speed, acceleration and trajectory.

2.2.2 Preprocessing

Articulographic measurements require some preparation. The planning and setting up of the experimental milieu itself is important. This means, for example, establishing a connection between AG501 and the prompter, constructing text-files to get the right target words on the prompter, or writing the recording scripts (Figure 6). The experimental set-up, including preparing the prompt script and experimentation scripts used during the research for this dissertation, was developed by Dr. Johan Frid and Dr. Susanne Schötz (Swedish Research Council, 2010-1599, PI Susanne Schötz). Minor changes to the scripts, such as the formatting of the prompt text, or changing the time window, were made by the author in collaboration with Frid and Schötz. Both researchers were present as assistants during the pilot recording, though not during the corpus recording during which they had advisory roles.

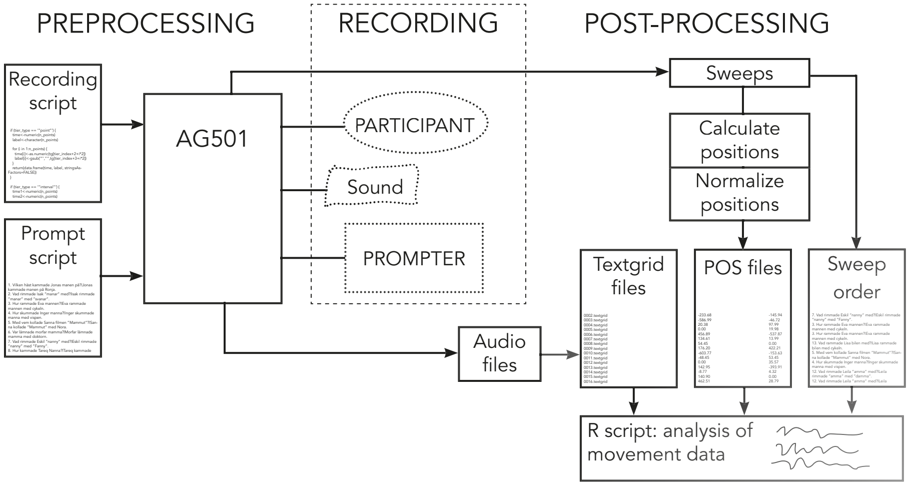


Figure 6. A flow chart describing the data collection processes
The procedure includes the preparatory work before recording, the actual recording, and the post-production of the data prior to the analysis work. Post-processing of the kinematic data takes place in the program of the AG501, while the audio files are processed in Praat. An R-script collects necessary information from the textgrid, POS and Sweep order files.

2.2.3 Recordings

The two recordings, made a few years apart, differed as regards speech material and speakers as well as the place of the sensors and some minor procedural details. On both occasions, sound and kinematic data were recorded simultaneously using a 3D ElectroMagnetic Articulograph (Carstens AG501), with an external condenser microphone (t.bone EM 9600) (see Figure 5). Furthermore, all sessions were recorded on video. This material was used for reference purposes during the period of analysis. For the same reason, photos were taken of the speakers to document where the sensors were placed. All participants were briefed about the procedure and signed a consent form.

2.2.3.1 Pilot study recordings

Pilot recordings of two speakers were made in December of 2014. The procedure followed a checklist for data collection, which included set up and pretesting (disinfecting sensors, testing scripts etc). When the participant arrived, she was given oral and written information on the procedure as well as a consent form. After metadata was collected the experimenter glued the sensors on the participant, who thereafter was placed in and hooked up with the AG501.

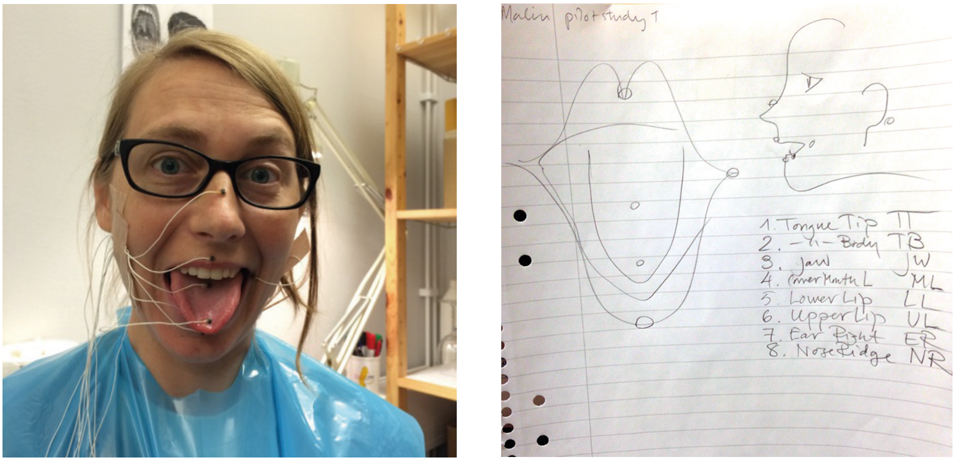


Figure 7. Picture of pilot study recording (December 2014).
The numbers on the image to the right (original drawing of the recording) indicate the order in which the sensors (from 8 to 1) were glued onto the participants.

Eight sensors were glued on the participants (see Figure 7): on the upper lip (UL) and the lower lip (LL) (both at the vermilion border in the sagittal plane), on the tongue body (TB), on the tongue tip (TT), on the jaw (JW), and at the left corner of the mouth (ML). The tongue body sensor, corresponding roughly to the tongue dorsum, was placed on the tongue where the participant made a bite mark after having stretched out the tongue as far as possible. Also, sensors were put on the ridge of the nose (NR) and behind the right ear (ER); the latter two sensors were used to correct for any head movements occurring during the recordings.

When the experiment started, the twelve sentences were shown, one by one in random order, on a screen for a few seconds and the participant was asked to read them out. The screen was placed 2-3 meters in front of the seated speaker, at eye level. Each sentence was recorded at one *sweep*. After the last of the twelve target sentences was shown the whole set was repeated. There were nine repetitions, in other words ten sets altogether, which amounted to approximately 120 recorded sweeps with articulatory and acoustic data for each speaker (240 sweeps in total). Sometimes a sensor fell off, which caused a pause in the recording, so that sensors could be re-glued, followed by a repeated reading of that word. Each recording took approximately 20 minutes (excluding time for set-up and post-processing). Post-processing (section 2.2.4) was done in similar fashion at both recordings.

2.2.3.2 Corpus recordings

Recordings for the corpus were performed during the spring of 2017 and were preceded by test recordings of two speakers in December 2016. In the test recordings, different test sentences were tried out, as mentioned in the previous section (e.g. clusters, nonsense words). At the same time, we also tested using sensors other than those in the pilot, such as sensors placed on the eyebrows and eyelids (for potential use in speech-gesture studies). The output of the test recordings is yet to be fully analyzed. In the end, we decided to use eight sensors in the corpus recordings, which are reviewed below.

The corpus speakers were again recorded in an ElectroMagnetic Articulograph (EMA), the Carstens AG501, with simultaneous sound recording. The data collection procedure followed the same checklist as in the pilot recording (and the test recordings). The eight sensors were placed at slightly different positions than in the pilot (see Figure 8). Data were collected from sensors placed on the midline of the upper (UL) and lower lips (LL) (at the vermillion border), on the jaw (JW), but also from three sensors placed on the midline of the tongue body. One sensor, corresponding roughly to the tongue dorsum (TD), was placed on the tongue where the participant made a bite mark after having stretched out his or her tongue as far as possible. Another sensor, corresponding roughly to the tongue blade (TB), was placed between the sensor at the back and a sensor placed approximately one cm from the tongue tip (TT). Since the two tongue body sensors do not necessarily occupy the same position in all 21 cases, and correspond only roughly to the tongue blade and the tongue dorsum, they are sometimes referred to as tongue body one, TB1, and tongue body two, TB2, (the front sensor being TB1 and the back TB2). These positions are found in posters or other research outcomes related to the dissertation project. In the written papers included in this dissertation, only one tongue body sensor placement is referred to (TB), which corresponds to the sensor furthest back, the tongue dorsum (TD). Since the AG501 collects data in five dimensions from each sensor, only one sensor is needed for head correction. However, to be on the safe side, we used two sensors for head correction, one behind the right ear (ER), and one on the nose ridge (NR).

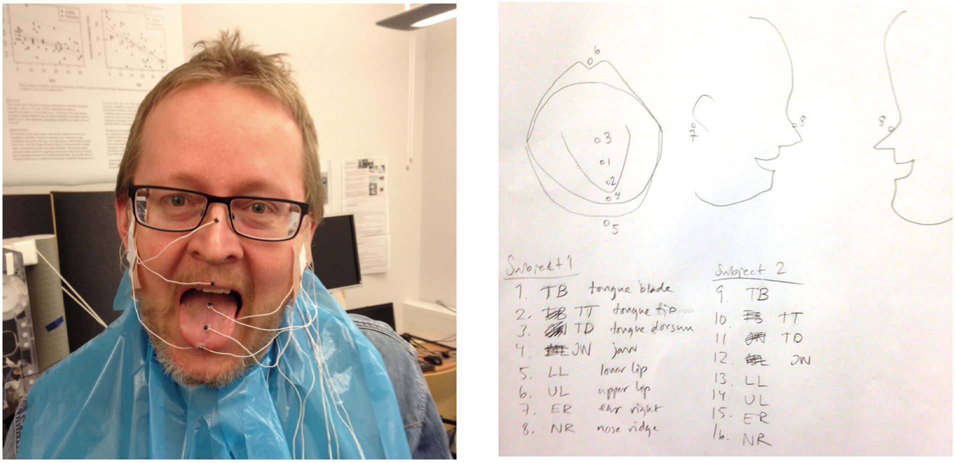


Figure 8. Picture of corpus recording (spring 2017).
The numbers on the image to the right (original drawing of the recording) indicate the order in which the sensors (from 8 to 1) were glued onto the participants.

Each corpus recording took longer time than the pilot, about 40 minutes for each speaker. The speech material was again presented on the prompter, both the leading question and the target sentence, and the speaker was instructed to read both sentences aloud. The 18 sentence sets were presented to each speaker in random order eight times (seven repetitions). This averaged 150 recorded sweeps per speaker, due to repetitions caused by sensors falling off, i.e. approximately 3000 sweeps in total. In addition, the first target sweep was preceded by a longer sweep (40 seconds) where the speakers spoke freely from an image displayed on the screen. These "free speech" sweeps have not yet been analysed.

2.2.4 Post-processing

After the recordings post-processing of the data was performed in the AG501. The post-processes were all identical (see Figure 6). Each sweep resulted in the creation of an *amp* file, which contains amplitudes of each channel (i.e. each sensor used) based on information from the nine transmitters. The CalcPos programme, included in the AG501, is then used to calculate positions for each sample, based on information in the amp files (Figure 6). This results in so-called *rawpos* files which contain information about each sensor's position in the five dimensions every 4 ms.

Speakers move their heads as they speak. Although head movements are of scientific interest, it is important to be able to isolate the articulators and thus eliminate the head movements. This was done next in the AG501 via the NormPos programme, where the references held the same position while the positions of the other sensors were rotated and changed (Figure 6). This resulted in a *pos* file for each sweep. The recordings also included audio recordings, which resulted in audio files. Next, the pos files and the audio files were transferred to another computer for further analysis, along with a text file containing the sweep order of each recording (see Figure 6).

Analysis of the kinematic data was done in R (R Core Team, 2015), which was obtained from the files using an R script.[7] However, in order to know what articulatory data was obtained, time indications were also required. These time windows were specified using input obtained from Praat textgrid files (Boersma & Weenink, 2018) (Figure 6). Thus, the time windows were based on segmentation and labelling of the sound files of the recording. The annotation of the sound files was made by the author, resulting in segmentation of consonants and vowels in the target words (C1, V1, C2, V2, C3).

The studies in this dissertation have focused on voiced consonants, especially nasals. Acoustically, nasals are complex because the sound travels through the nasal cavity. The lowering of the velum, which allows for this changed direction of sound, causes the formation of anti-formants. In combination with the closed mouth it in turn makes the nasal sounds quite low. However, the formant transitions between vowels and nasals are relatively clear, especially where word-initial nasals are involved, which made segmentation fairly straightforward. The approximants, on the other hand, were more challenging to segment. The formants and the intensity did not assist as much as when judging the nasals; however, listening carefully to the material helped make reliable assessments.

As mentioned, the kinematic data was designed to be collected from the pos files using the segment information of the textgrids. For example, the command '*look for lip movements during V1*' collected position data from the sensors on the upper and lower lips of the stressed vowel. In the script, not only time was specified, but also the nature of the movement: '*look for the <u>fastest</u> lip movement during V1*' collected maximum velocity of the lip sensors' movements in the stressed vowel. These commands were revised for each study depending on the purpose of the research (for more information, see each of the measurements sections below, 2.3ff).

The commands in the script also needed to be slightly revised for each speaker because of the speaker-dependent movements of the articulators. This was done, for example,

[7] The R script used to collect the kinematic data was written and developed by Dr. Johan Frid and Dr. Susanne Schötz, and revised by Dr. Johan Frid and the author. Minor adjustments to the script, including specifications of speakers, sensors, measurements or target words, have been made by the author.

by letting the time window start a little earlier: '*look for lip movements* <u>*20 ms before start*</u> *of V1'*. Speaker-dependent time windows were a problem mainly for the tongue body data, which varied both between and within speakers. That is, similar to "noisy" data, as seen in other studies (Cho, 2002; Gao, 2008). Since there were many speakers, the time windows needed frequent adjustment. Assessments of what adjustments were required was done by the author by plotting the kinematic trajectories for visual inspection. The script also included some elementary calculations used in the studies, such as calculating *lip aperture* in R as the Euclidean distance between UL and LL. The *tangential velocity* was also calculated in the script. These calculations and equations are specified in the measurement sections where applicable (2.3ff).

First derivatives, i.e. velocity, have already been mentioned as a specification of how a movement was measured. When it comes to articulation one should keep in mind that the articulators should not collide. Therefore, it is a significant event when velocity changes, as an adaptation to the oral cavity or/and to the different targets. In this dissertation, second derivatives, i.e. acceleration, have also been investigated. Thus, one possible command might be: *'look for when lip movements* <u>*change velocity most*</u> *during V1'*.

Furthermore, in order to be able to extract landmarks from second derivative (acceleration), the position data needs to be filtered and smoothed. We have filtered the signal with low-pass filter, using the R function loess (span = 0.1). The fitting is done locally, which means no time delay. The acceleration signal is thus very simplified; we are aware that RMS (Root Mean Square) values increase as smoothing increases which can be a problem for the fast movements (Hoole & Xierdt, 2010; Hoole, 2014).[8] But, on the one hand, we do not measure the fastest movements of the oral cavity. Furthermore, filtering out disturbing signals is a way to move this line of research forward.

2.3 Measurements

The following section reviews the technical and the procedural details of the measurements used in each of the four papers. An evaluation of some of these measurements is made in Chapter 4 (section 4.3). The four studies consist of acoustic or articulatory measurements, or both, as specified in each section. Moreover, the studies are based on different parts of the recorded material: Paper 1 is partly based on a subset of the pilot material; Paper 2 and 3 on a subset of the corpus data; Paper 4 on yet another subset of the corpus data. Which subsets is indicated in the following

[8] RMS value is a measure estimating distortion of the data.

sections for the sake of clarity. The statistical analyses are also sometimes mentioned where clarification is needed, but the overall statistical analyses are otherwise addressed in section 2.4.

2.3.1 Paper 1 – Acoustic and articulatory measurements

Paper 1 includes both acoustic and articulatory measurements of data from two different experiments with different sets of speakers. Experiment 1 comprises acoustic data from material supplied by 12 South Swedish speakers. The data includes the target words /ˈbiːlɛn/ (*the car*) and /ˈbiːlar/ (*cars*) produced with high prominence and with different discourse context conditions. This material is from another research project (Swedish Research Council, 2009-1566, PI Johan Frid). Experiment 2 in Paper 1 uses both acoustic and articulatory data from a subset of the pilot project data (which included two speakers). Identical target words are used to match Experiment 1: /ˈbiːlɛn/ and /ˈbiːlar/.

2.3.1.1 Acoustic measurements

Both of the data subsets were manually segmented into consonants and vowels (the first three segments: C1, V1 and C2). The f_0 data from Experiment 1 was manually corrected for pulses. Both of these procedures were done in Praat using ProsodyPro (Xu, 2013). While the f_0 data outcome was only visually analysed the segmentations were also used to make acoustic measurements (see Figure 9). These were:

- Segment durations of the first three segments /bi:l/ (C1, V1 and C2). Collected using ProsodyPro in Praat. Data from Experiments 1 and 2.
- Formant measurements: mean F_2 and F_3 from the occlusion phase of /b/ and at onset of /i:/. Collected in Praat. Data from Experiment 2.
- Voice quality annotations. An annotator decided on presence/absence of creaky voice in /i:/ and /l/. Data from Experiment 1.

2.3.1.2 Articulatory measurements

The articulatory measurements in Paper 1 were all on the data from Experiment 2. The measurements focused on the timing of tongue positions and on the openness of the lips during the word-initial segments /bi:l/ (see Figure 9). After the data collection and the post-processing, the following preparations were made for measurement: in R we located the target words using the textgrids from the already segmented speech. Having thus created a time window, specifications were required to retrieve the correct articulation data. We chose to collect the vertical kinematic data from the sensors on the lips (UL, LL), the tongue body (TB) and the tongue tip (TT) (note that the TB data is from the sensor referred to as TD; see section 2.2.3.2).

First, we needed to pin down the point in time when the lips are opened or closed. Thus, we calculated lip aperture as the Euclidean distance between UL and LL:

$$Lip\ Aperture = \sqrt{(ULx - LLx)^2 + (ULy - LLy)^2 + (ULz - LLz)^2}$$

Two lip positions were of interest here: the start of the consonantal gesture movement towards /b/ and how long time it took before it reached its target (see Figure 9). The start was set as a data point with maximum openness during the anticipatory lip movements. This occured even before /b/ so the time window had to be adjusted to include these anticipatory movements. The target was set as the minimum lip aperture when the lips are closed during /b/.

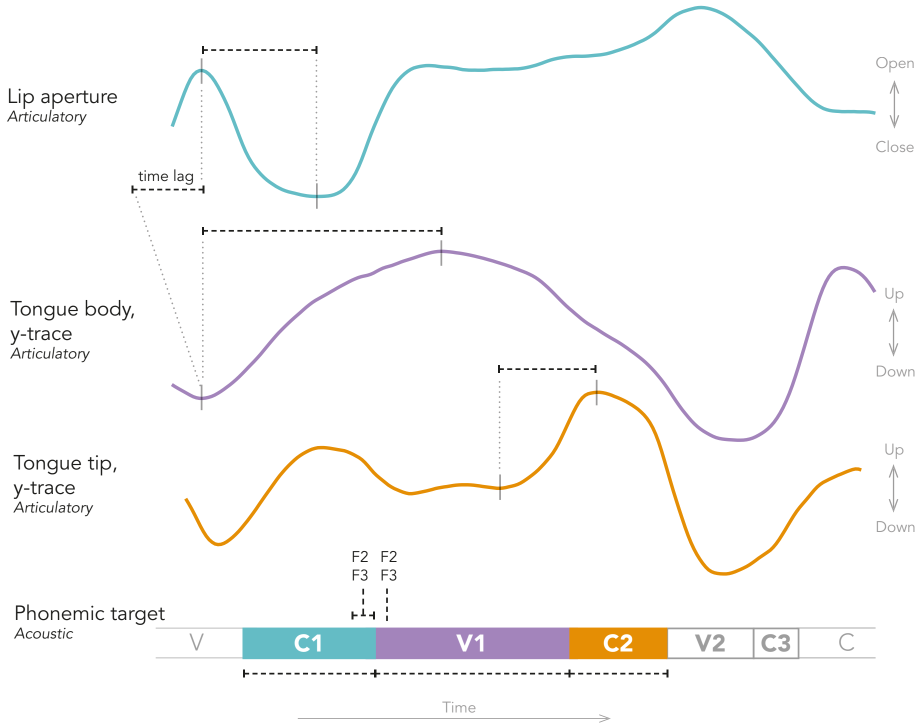


Figure 9. Measurements in Paper 1.
The six articulation data points (grey vertical solid marks) with the four measurements (black horizontal dashed lines) in Paper 1: time lag between C1 and V1 onsets; lip aperture closing interval of C1; tongue body rising interval of V1; and tongue tip rising interval of C2. The acoustic measurements included: duration of C1, V1 and C2; and formant measures (F_2 and F_3) on the occlusion phase of C1 and at a time point in V1 onset. In addition, not in this figure, annotation of creaky voice in V1 and C2, and visual inspection of f_o.

Tongue body data was collected in a similar fashion. However, instead of a calculated measure as in lip aperture we used the actual trace of the TB sensor in vertical angle (y-trace). We were interested in the start of the tongue body rising and the time it took for it to reach its target, the constriction in /i/. Since the tongue body starts to rise during or even before the adjacent segment, the time window for collection of the minimum position of TB was specified to occur during or before /b/. The target of the gesture was located as the following maximum position of the TB sensor.

To collect data on the second consonant (C2) in the target word, /l/, we again used the trace of an actual sensor, but this time the tongue tip (TT). The data points collected were the lowest position (minimum) at the end of the vowel segment, and the following highest position (maximum) of the TT sensor. The procedure was similar to the procedure on the tongue body in that the adjacent segment was used to locate the minimum.

The two data points that signalled the start of the bilabial movement and the tongue body were also used to calculate onset time lags. This meant the points: maximum opening of the lips and minimum position of the tongue. Time lags were calculated as the distance between them in ms.

2.3.2 Paper 2 – Articulatory and acoustic measurements

The measurements in Paper 2 focused on the articulation of the lips and the tongue, but the acoustic f_0 signal was also addressed. A subset of the data from the corpus was used, which included words with the word-initial segments /ma/. This meant that the target words had different stressed vowel lengths, as well as different word-medial consonants and different word endings. A total of four A1 – A2 pairs were used, i.e. eight target words in total (see Paper 2 for further information).

2.3.2.1 Articulatory measurements

The articulatory measurements were made on the lips and on the tongue body. Therefore, data were obtained from the sensors UL, LL and TD. As in the previous study, lip aperture was calculated in R (see section 2.3.1.2). In other words, all lip measurements concern the distance between the lips, i.e. the lips were treated as a common articulator and not as two separate ones. Data from the TD sensor was also pre-processed for further measurements by calculating the tangential velocity (Löfqvist & Gracco, 1999):

$$Tangential\ velocity = \sqrt{(\dot{x}^2 + \dot{y}^2)}$$

By "tangential" is meant that the vertical and horizontal movements are treated jointly, not separately. Thus, several of the subsequent temporal and spatial measurements are calculated on both the tongue's lowering and its retraction simultaneously. Measurements were also made of the vertical movement of the tongue alone, that is, its lowering. The choice of measurements was based partly on previous studies (for details, see Paper 2, section 1.5), and partly on the basis of observing the data. The measurements can be divided into three main areas: *coordination*, *temporal* and *spatial*.

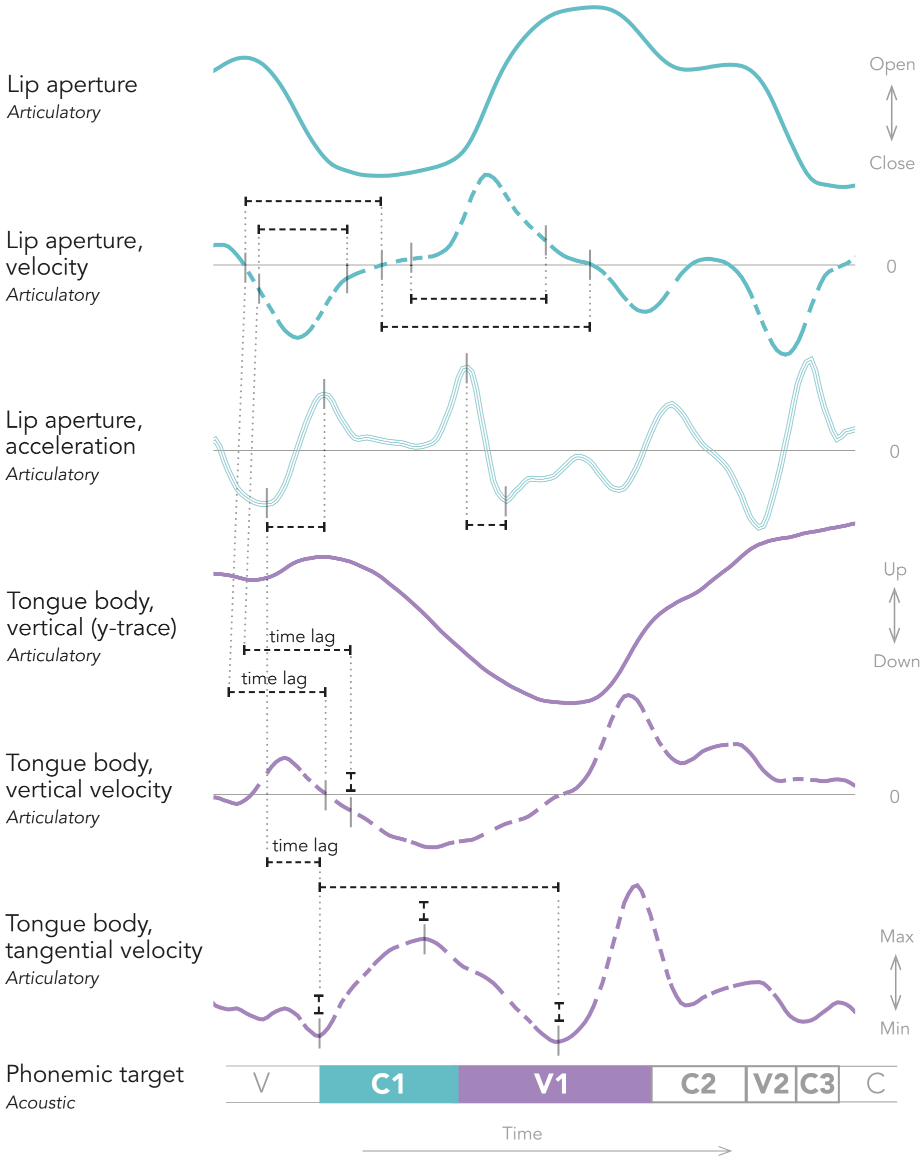


Figure 10. Measurements in Paper 2.
The fourteen articulatory measurements that comprised Paper 2. Three intervals are time lag measurements (*coordination*) between the starts of the lips and tongue body movements. A total of six intervals (*temporal*) are measured as to the movements of the lips; three at the closure and three at the opening of C1. One interval (*temporal*) measures the length of the tongue lowering. Four of the measurements (*spatial*) are of the height of the tongue at different time points.

The *coordination* measurements in the study consist of three CV time lag measurements. These measure how well C onset and V onset are timed, and the three measurements were completely replicated from previous studies. Figure 10 shows in detail what the different time lags measure. Firstly, we have time lags between the vertical tongue body movement and the velocity of the lips. There are two different onsets here: either zero-crossing velocity, or 20% threshold from zero to peak velocity (local minimum). Thus, these two onsets apply to both articulators.

The third time lag measurement refers to the distance between the acceleration of the lips and the onset of the tangential tongue body movement. Thus, for the lips, it is the second derivative of lip aperture, and the landmark is maximal acceleration (local minimum).[9] Tongue onset is placed at minimum tangential velocity. All three time lags mentioned are measured in ms, and positive values mean that the tongue onset follows the lips. In case of negative values, the tongue has started its movement towards the open vowel before the lips begin to close.

The *temporal* measurements arose partly as a way of evaluating the CV time lags, partly as a result of discoveries made in Paper 1, where A2 implied a longer tongue body movement than in A1. Thus, lip measurements were based on the onsets used in the three time lags, which meant lip closure interval based on: zero velocity onset and target; 20% threshold to peak velocity (local minimum) onset and target; and maximal acceleration onset (local minimum) and maximal deceleration target (local maximum) (see Figure 10). Similar intervals were measured for the lips opening of /m/, except for 20% threshold to peak velocity, which was now a local maximum; and, maximal acceleration which was instead local maximum, while maximal deceleration was now local minimum. Since there were many landmarks (12 time points), the collection of these measuring points caused necessary, and time-consuming, manual adjustments of the time windows for each speaker (as mentioned earlier in section 2.2.4, this chapter).

The measurement of the length of the tongue lowering (the palatal wide interval) was made between the vocalic onset and the target, both of which were minimum tangential velocity (see Figure 10). Thus, the same onset landmark was used as in one of the coordination measurements above. All intervals were measured in ms, and any negative values were sorted out before the statistical analyses (because if they existed, they were due to collection errors).

In this study, *spatial* measurements were made only on the tongue body. The focus was early on the spatial position of the tongue, because previous studies have shown different tongue body heights between different tones. Thus, one of the spatial measurements is a replication of a previous study: vertical tongue body height at 20%

[9] *Local minimum* refers to the mathematical concept, i.e. when the curve goes below the 0-line (*local maximum* when there is a positive value). However, in the dissertation I mostly use the terms *maximal acceleration* and *maximal deceleration* (which can be interpreted more qualitatively).

threshold to peak (vertical) velocity (local minimum). Incidentally, it is the same landmark used by one of the three CV time lags. Another of the four spatial measurements is based on, but not a replication of, yet another study: vertical tongue body height at minimum tangential velocity at vocalic target (see details in Paper 2; sections 1.5 and 2.4). The two remaining tongue body heights are also vertical heights: at minimum tangential velocity at onset, and at maximum tangential velocity (can also be referred to as peak tangential velocity). Together with the tangential onset and target, these three spatial measurements of the tangential velocity provide a relatively representative view of the tongue body movement over the course of the word-initial segments. The tangential onset and target from which vertical height is taken are also the same landmarks used at the temporal tongue body interval (Figure 10).

2.3.2.2 Acoustic measurements

The acoustic measurements in Paper 2 consisted only of f_0 measurements. In the study, the f_0 differences between the word accents were handled as two tonal categories. Therefore, qualitative assessments of the tonal measurements were made to establish that the word accents in the material actually followed the expected tonal pattern. However, most of the articulatory measurements occurred in the vicinity of the word-initial consonant and therefore it was important to establish that f_0 varied during that time, and not only during the vowel segment. The qualitative assessments were therefore supplemented with statistical analysis of pitch during the C1 segment.

For the qualitative assessment and the statistical analysis, separate preparations were made of the f_0 data. In both analyses, time normalizations were made (see details in Paper 2). Prior to the statistical analysis, a normalization based on words was also made, while the qualitative assessment included a speaker normalization. In addition, preparations for the qualitative assessment included a manual correction of the data (using ProsodyPro; Xu, 2013). The statistical analysis consisted of a t-test (see further information in section 2.4).

2.3.3 Paper 3 – Articulatory measurements

In Paper 3, the same subsets of the data were used as in Paper 2, i.e. word-accent pairs starting with the CV sequence /ma/. However, this study contains only articulatory measurements, and only on data collected from sensor JW, i.e. the jaw, or mandible, movements. In turn, both a temporal and a spatial analysis of the jaw were made.

The temporal measurements were made on the second derivative of the jaw's combined vertical and horizontal movements (Figure 11). Thus, tangential velocity (Löfqvist & Gracco, 1999) ($[v = \sqrt{(\dot{x}^2 + \dot{y}^2)}]$) was first calculated for the jaw movement before the second derivate (acceleration, or *tangential acceleration* to be more precise) were collected on the data. As already mentioned in the previous section (2.2.4), position data were smoothed to obtain a less fluctuating acceleration signal.

All kinematic trajectories (y-trace, x-trace, velocity and acceleration) were plotted for visual inspection by the author (similar to Figure 11). Only then were landmarks marked for further calculation of intervals that could be of interest, i.e. to capture the jaw movements. These landmarks were the maximal acceleration (local maximum) and maximal deceleration (local minimum) of the jaw opening and of the jaw closing, i.e. a total of four timing landmarks (Figure 11).

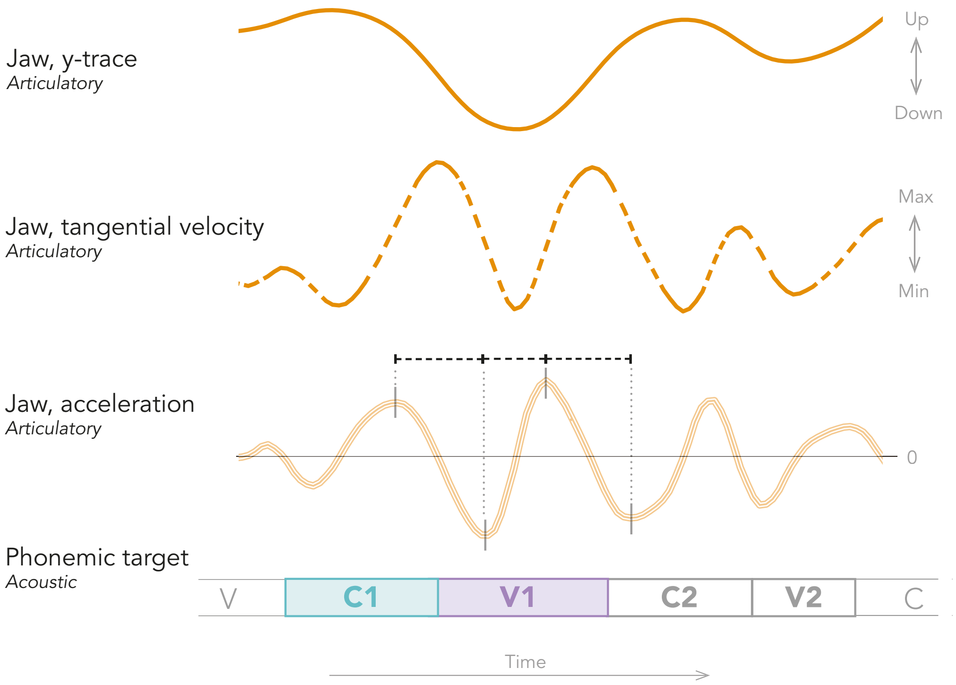


Figure 11. Measurements in Paper 3.
The four landmarks and the three intervals that were measured in Paper 3.

The distance between the two constituting either the opening or the closing of the jaw was calculated, and they were labelled *jaw opening* and *jaw closing*. A third interval between these two intervals, i.e. between maximal deceleration (local minimum) of the jaw closing and maximal acceleration (local maximum) of the jaw opening, was measured and named *jaw open posture*. Hence, this third interval is the time that the jaw stays open without any rapid changes, and thus presumably, with less active movement (Figure 11). The intervals were each measured as duration in ms. The three intervals are also measured jointly in what is referred to in Paper 3 as *Total jaw duration*.

Paper 3 also included measurements on the spatial data, but these were only qualitatively assessed, i.e. no direct spatial measurements were made. For the qualitative analysis, only the vertical data was used, and the analysis was prepared as follows: first, the target words were normalized by time. Then followed a comparison between two different ways to normalize the vertical data, either on each word or each speaker. Plotted regression lines visualized the differences of vertical position (see Paper 3).

2.3.4 Paper 4 – Acoustic measurements

Paper 4 is an acoustic study dealing exclusively with segment duration. In this study, no comparisons are made between word accents, and the independent variables are instead different features of the word-initial (C1) and the word-medial (C2) segments. Since we therefore wanted to control f_0, only A2 words were examined. Thus, the study is done on a subset of the corpus data containing A2 words where the vowels are the same but where the consonants differ: /C1aC2a(C)/ (however, the stressed vowel can still be both long and short).

Since we do not use the articulatory data of the corpus, all measurements are made instead of the information received in the Praat text grid files. That is, the segmentation that was previously used as a time window to find the right articulation in time, is now used in its own right. Thus, the dependent variables are duration of different consonant segments: word-medial /n:/, /m:/ and /l/; as well as word-initial /m/ and /n/. The duration of C1 and C2 is calculated as the distance from the onset to the offset mark (Figure 12).

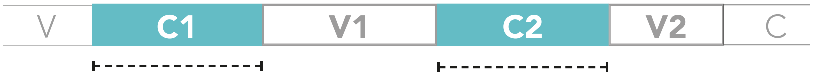


Figure 12. Measurements in Paper 4.
Paper 4 includes acoustic segment duration of the word-initial (C1) and the word-medial (C2) consonant. Dashed lines indicate what is measured.

The difficulties met with in segmentation were of varying degrees, as already mentioned in the previous section 2.2. While segmentation of nasal stops was relatively straightforward, the word-medial continuant /l/ was a little more problematic. This is further addressed in Paper 4. There was also a certain difference between the clear word-initial segments, and the word-medial segments which sometimes required a little more analysis before labelling. The segmented stressed vowel (V1) was also used in the study but not for the same purpose: raw duration data on C1, V1 and C2 were presented for further interpretation.

2.4 Statistics and analysis

The following section presents the overall statistical methodology, together with a brief summary of each study. Further details about the different statistical approaches in each study are found in the individual articles. In Paper 1, however, information on the statistical methods is sparse, mainly due to limitations of space. Therefore, slightly more details on the statistics are presented here to make up for the lack of it in the paper.

The inferential statistics in the dissertation mostly focuses on generalized linear mixed effects regression models (GLMMs). These were all done in R using the lme4-package (Bates et al., 2015), and the lmerTest-package (Kuznetsova et al., 2017). All of the studies were exploratory within-subject studies with categorical variables. In the corpus there are eight blocks of repeated target sentences (seven repetitions). This means that the results include repeated measures. Most of the measurements centre around time as a concept of articulation, either as duration of the movement or as in timing with other articulators or other segments. f_0 data stands out as mainly being analysed by qualitative inspections, although in Paper 2 a Welch Two Sample t-test was performed. Qualitative inspections were also used for the spatial positions on the tongue body and on the spatial jaw data. All inferential and descriptive statistics as well as graphs were performed in R (R Core Team, 2015). For pedagogical reasons visualizations of the results have sometimes been done in InDesign and Illustrator (Adobe Inc., 2019).

In Paper 1, which included material from two different data sets, the mixed models were performed on several types of measures. The voice quality data included annotation judgements from one dataset, which was on the 12 speakers from the project Function- and production-based modelling of Swedish prosody (Swedish Research Council, 2009-1566, PI Johan Frid). The duration results included material from the data subsets of the two experiments. The results based on data from the larger group of speakers were implemented using GLMMs (specified for clarity in Table 1, which summarizes fixed and random effects). The results that were based on data from the pilot study, i.e. only two speakers, were achieved as follows: a standard regression model was used for the durations and formant data, as well as the articulatory measurements. There, the effects of the independent variables *word accents* and *speakers* were assessed.

In Paper 2 and 3, the focus was on articulatory measures (apart from the f_0 data using the t-test in Paper 2) along with more speakers (19). The decision was made to compare models using Akaike Information Criterion (AIC). The more complicated model was chosen only when it exhibited a lower AIC value of at least 2 (following Wieling & Tiede, 2017). As a result, *vowel length* was only added in models if it was warranted (Table 1). Furthermore, in Paper 2, most models on the measurements did not use *speaker* as random effects slope, because the model without got the same (or an even better) result.

Table 1. Mixed effects regression models

Summary of the GLMM models used in the four papers. **Factors included only if complexity was warranted (according to AIC).*

PAPER	MIXED EFFECTS REGRESSION MODELS		
		FIXED EFFECTS	RANDOM EFFECTS
1	Voice quality results	Word accent Gender	Speaker (random intercepts) Discourse context (random intercepts)
	Duration results	Word accent	Speaker (random intercepts) Discourse context (random intercepts)
2	All results	Word accent **Vowel length*	Speaker (random intercepts, **random slopes*) Word (random intercepts)
3	All results	Word accent **Vowel length*	Speaker (random intercepts, random slopes) Word (random intercepts)
4	All results	Word	Speaker (random intercepts, **random slopes*)

In Paper 3 *speaker* was instead specified as both random slope and intercept. This approach seemed appropriate because an initial analysis showed that a model with *speaker* was warranted, and because all the articulatory data was from the same articulator – the jaw. Furthermore, in both Paper 2 and 3, which complement each other in many ways, spatial positions on data from the articulators were analysed by qualitative inspections by the authors. These were time-normalized vertical positions and because of the normalizations it seemed not fruitful to do further statistical analysis on those particular calculations.

No articulatory measurements were made in Paper 4, however GLMMs were still used. The duration measurements were transformed to a log scale (log base 2 in R). The models used for the log data on C1 and C2 used *speaker* as random effect with random intercept (random slope only added after model comparisons using AIC) (Table 1).

3 The studies

This chapter contains summaries of the four studies. Each summary is followed by a short general discussion, which also shows whether the dissertation's general research questions are answered.

3.1 Paper 1: Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents

3.1.1 Summary

Paper 1 is a conference paper and the first of three papers on the Swedish word accents. The study is an explorative investigation into the multidimensional nature of the word accents. It addresses specifically the different tonal patterns in A1 and A2 as spoken by South Swedish speakers. In Paper 1, all target words are highly prominent (referred to as *focal accent* or *sentence-level intonation* in the paper) and some also comprise different discourse context conditions. The predictions were that we should be able to find both acoustic and articulatory evidence of the different word accents, other than those of the more commonly established f_0.

The acoustic parameters investigated were – beside f_0 – voice quality, segment duration and formants. For instance, we predicted duration differences based on the timing of the f_0 excursion, that is, the stressed vowel should be of equal length but the following consonant longer in A2. As for articulatory evidence, we expected to find signs of a c-center effect, meaning an onset time lag (referred to in the paper as *gestural overlap*) between when the articulation of consonant and vowel, respectively, started. Based on previous modelling of tone gestures, the time lags might differ in length between the word accents.

The results led to the conclusion that the acoustic duration of the post-vocalic consonant was indeed longer in A2. This was explained by the tonal fall during the post-syllabic segments, in contrast to no tonal fall in A1. Although the acoustic segment durations did not differ otherwise, we found other evidence of a link between tonal context and the underlying segments. For instance, the articulation of the tongue body was noticeably reaching its target (that is, the highest tongue position) earlier in A1 than in A2. Similarly, creaky voice was more present in A1 than in A2, and in addition already during the stressed vowel in A1. These patterns, we assumed, were explained by the early peak and consequently early fall in A1. Besides concluding that the word accents in the study did portray the different f_0 patterns mentioned, a rather unrelated and unexpected result was found: the f_0 pattern in one of the sentence-level (discourse context) conditions deviated from that of the others. The condition "exclamation" displayed an extra tonal peak following the first one. This pattern is known as a *two-peak* pattern and is common in other Swedish dialects, such as Stockholm Swedish, and normally not found in the South Swedish phonology.

Other results of the study seemed to be speaker-dependent, meaning that, according to our measurements, speakers articulate differently between A1 and A2 but not in the same systematic way. For example, one speaker revealed a longer time lag in A2 accompanied with differing formant changes at vowel onset between the accents while the other speaker demonstrated a shorter lip movement in A2 along with differences in formants during the occlusion in /b/, as well as an effect by word accent on the tongue tip movement in the post-vocalic consonant /l/.

3.1.2 General discussion

The results of the speaker-dependent behaviours may be interpreted: articulation does not always occur in a systematic way. On the other hand, the way we measured may not have captured the systemization. We are so used to the acoustic parameters and the way we look at speech as in segments and phonemic targets that we sometimes do not attach importance to the variation of those established parameters. If the variation is not large enough to function perceptually, they can easily be ignored. However, very small variations in an acoustic pattern may indicate a significant underlying connection in the articulation.

The same approach should be applied to articulatory parameters. The kinematic measurements we used in this study, such as time lags or interval duration, may not have captured the speaker variation satisfactorily. Although Articulatory Phonology and other theories/models have developed some fruitful measures, there is still a broad approach in how to measure articulatory movements. Furthermore, during the work in Paper 1 it was assumed that the c-center effect arises due to delays in onset time. However, that connection has later appeared increasingly difficult to interpret. This study enabled me to understand the spectrum of established and new methodology, and which of all the myriad paths provide us with the most useful information.

This study was mainly focussed on clarifying how the tone gestures in Swedish are executed. The idea was to be guided by the acoustic analyses and the articulatory coordination. However, the question is more complex, and the study was not able to say what a tone gesture is. It rather helped to put the finger on the question. The most important contribution of Paper 1 is its multi-dimensional aspect. It shows the importance of linking articulation with acoustics, and of the consequence of simultaneous f_0. The results show that several parameters are influenced by f_0, which indicates that f_0 is evidently intertwined with the rest of the articulation. Moreover, we cannot understand timing and time aspects at the segmental level without linking it to how the articulation is coordinated. Articulation is dynamic and this means that a change somewhere affects the other parts of the complicated design of the oral cavity. It follows from this insight that the next study ought to be a deep dive into articulation.

3.2 Paper 2: Word-initial consonant-vowel coordination in a lexical pitch-accent language

3.2.1 Summary

Paper 2 is an extensive journal article which is primarily about articulation. It is the first paper in this dissertation that makes use of a subset of the corpus material. Similar to Paper 1, the study is an explorative investigation into the SWAs, and it develops some of the results of the previous article. Paper 2 deviates from Paper 1 in that it concentrates solely on, and goes deeper into, the word-initial consonant (C), the subsequent vowel (V) and the overlapping coordinative relationship between them (referred to in the paper as *CV coordination*). This includes addressing the time lags between C and V onsets and how this has previously been measured, with or without contrastive tonal contexts. It also takes a more thorough approach to the lips and the tongue body, measuring both the lip closing and the lip release, as well as duration and height of the tongue body movement. In total, the study includes fourteen measurements: ten temporal and four spatial. These measurements feature first and second derivative calculations of vertical and tangential movements. That is, there are no measures based on the actual x- or y-traces of the sensors, contrary to Paper 1. Needless to say, the measurements in Paper 2 are more complex and in a wider sense more methodically based on previous research.

Like Paper 1 the study focuses on a specific CV sequence, but this time /ma/ (in Paper 1 it was /bi/). Furthermore, the subset of the data from the corpus comprises target words with different vowel lengths (in Swedish this entails different vowel quality as well) and different word endings (see Section 2.1.2). Whether these differences had effects on the results is addressed in the paper.

The main results show that of six measurements on lip movements the only one that differed systematically between A1 and A2, across speakers and words, was the second derivate interval of the closing of the lips. That is, the bilabial closure interval between maximal acceleration and maximal deceleration is shorter in A2 than in A1. The lip landmarks based on first derivate do not capture these systematized movements between the word accents; they only register a difference between the two vowel lengths (further discussed in section 4.1 in Paper 2). Nevertheless, this finding tells us that at word onset the position of the lips differs between the word accents.

The tongue body movements also differed between the SWA. In A2 the tongue body took longer to reach its target (with target meaning the point in time of minimum velocity). There was also a significant difference, across speakers and words, as regards spatial position when the tongue was moving at full speed. One possible interpretation of these results is that the tonal context affects velocity and thus also the position of the tongue body and its trajectory. Paper 2 includes a discussion of the correlation and possible mechanical coupling with the structures enabling f_o (Paper 2, section 4.2).

The results in Paper 2 also included measures of the *CV time lag*, meaning difference in time (ms) between the point when the lips start to close (C onset) and the tongue body starts to lower (V onset). In this study the CV time lag only became significant when the C onset was a second derivate lip landmark and the V onset a first derivate tangential landmark. In other words, between the point when the lips accelerated the most and the point before the tongue started to move both backwards and downwards. However, these movements were relatively synchronized and the time lag was scant.

To conclude, the articulation of the CV sequence /ma/ is indeed affected by the tonal context. Our study shows that both lips and tongue body move differently – separately and in accordance with each other – depending on the tonal context. We assume this is mainly due to a mechanical connection to f_o, but this is in turn largely dependent on when that connection is active, which may well have phonological connotations. The study is the first extensive articulatory study, as to measures and speakers, on the SWAs. Moreover, Paper 2 makes a case for including as many speakers as possible (since articulation is highly speaker-dependent) and encourages cross-linguistic studies that take a full dynamic account of the articulatory movements (e.g. acceleration, velocity and position).

3.2.2 General discussion

Paper 2 is the first study in the dissertation that looks seriously at what is systematized in the articulation. Furthermore, it also tries to answer both how the different tonal contexts affect the articulatory movements at word onset, and how the connection between the consonant and the vowel in particular is affected. Over time, it also turned out to be a study of what constitutes an articulatory gesture.

Paper 2 is an extensive study which has its advantages and disadvantages. Since previous articulatory research is relatively unanimous about what should be measured but not necessarily HOW it should be measured, we felt compelled to use fourteen different measurement points. Still, it seemed important to include differences in ways of measuring CV time lags. The time lag measurement is becoming more common in phonetic research and an evaluation of HOW it is measured is needed. It is often used as a phonological measure. It is still uncertain if this is a reasonable approach.

The study pinpoints the need for many speakers in articulatory research. Two or five speakers are not enough when studying articulation because variation is large, especially in the movements of the tongue body. In this study, the vertical measurements varied a great deal, which indicates the need to include horizontal movements. In this particular case, however, it is possible that since we did not use a bite plane during the recordings, the vertical measurements were affected by the main angle and therefore did not reach significance. This, in turn, could have affected our results on CV time lags based on vertical velocity.

Results show that lip movements differ between word accents, either when the movement starts or when it ends, or perhaps as a total speed difference. A subsequent study is in progress and preliminary results show that the lip positions do indeed differ between the different tonal contexts. Further research on the relationship between the position, velocity and acceleration is needed to map what is actually happening here. Although the lips appear to be positioned differently between the two tonal conditions, it is not clear how this relates to the other articulators. For example, the link to the jaw must be established, that is, if it lowers more for A1 (this is addressed in Paper 3). The tongue body is an articulator with many degrees of freedom and is therefore articulatory complex. It also functions in relation to the jaw movement, meaning probably that position and speed are affected by how the jaw moves. At present, however, we cannot discern what is due to the jaw and what the individual movement of the tongue is because, in this study, the tongue body data used is not separated from the jaw.

3.3 Paper 3: Jaw movements in two tonal contexts

3.3.1 Summary

Paper 3 is a conference paper and serves partly as a complement to Paper 2. It examines the effect of tone context on jaw articulation. Accordingly, it examines the role of the jaw in the same subset of data as in Paper 2. The study can also be seen as an exploration of the timing aspect of acceleration (second derivative) of a movement. Therefore, Paper 3 also stands on its own as an individual study. The same subset of the data is used as in Paper 2; hence, the target words consist of A1 and A2 words beginning with the CV sequence /ma/ with different vowel lengths (which also entails a different vowel quality), and different word endings. The research question that instigated the study was whether we would find differences in jaw movements between the two tonal contexts. Because of what we know about mass-spring systems and articulatory dynamics, we were interested in finding an onset and an offset of the movement that was due to acceleration. During the jaw cycle of opening and closing we found two points of maximal acceleration and two of maximal deceleration. Three well-defined intervals then emerged which were labelled: *jaw opening*, *jaw open posture* and *jaw closing*. The analysis consisted of duration measurements of the intervals as well as position measurements of the vertical height of the JW sensor.

Results showed that the jaw opening and jaw opening posture were affected by the tonal context. In the A2 words, in which the tone started low and then rose, there were longer jaw opening intervals than in A1. The same pattern was found at jaw open posture but only for the words with long vowels: thus, the jaw stayed open longer for the A2 words when the tonal peak was within the boundaries of the vowel. The main findings also included a significant difference between the vowel lengths during the jaw opening posture interval. The long vowels had longer intervals (not surprisingly), and in addition varied more in length. Interestingly, there was no difference between vowel lengths word-initially at the jaw opening interval.

In Paper 3 we suggest that the different intervals interact with the timing of the tones. The fact that the jaw opening is shorter in A1 may have to do with a truncated jaw movement due to the early tonal peak. In the same way, the extended open posture could be an adaptation to a late tone peak in A2. The total jaw duration was longer in A2, which may support this interpretation and, moreover, follows previous research on extended syllables in A2.

Unfortunately, the results were inconclusive as to the spatial position. The vertical height data was normalized partly by word and partly by speaker. The two different approaches yielded disparate results which were difficult to interpret. Likewise, we did not get any clear results on the jaw closing interval, probably due to the different word endings. We have previously found that certain, but not all, word-initial movements are influenced by the ending of the words, a discovery that inspired Paper 4.

3.3.2 General discussion

The purpose of Paper 3 was to investigate the articulation of the jaw and how it might be affected by the tonal contexts of A1 and A2, in addition, to investigate whether the acceleration and deceleration can have a phonological role in the syllable. The choice to focus on acceleration of the jaw was initially due to avoiding the motion plateaus that arise from velocity differences. We had seen from the lip movements that they could have minimal movement for an extended period of time. We assumed that was the case with the jaw as well because of its similar degree of freedom as the lips, which means that its motion is less prone to constant dynamic movements as for example the tongue body.

Another reason for focusing on acceleration was its relation to velocity and position. In a damped mass-spring system, a body decelerates when the target is overshot, but in the closed oral cavity it also needs to decelerate in time so as not to collide with other articulators. Thus, it was assumed that any changes in acceleration or deceleration would suggest significant phonological positions of the jaw. This was perhaps an excessive and simplified assumption and the results became difficult to interpret in that context. However, the intervals detected and measured may prove to be important to further understand the timing of phonological differences. For example, if we had measured the jaw intervals by velocity, we might not have been able to detect the vowel length distinction. A longer section on this topic can be found in Chapter 4.

In Paper 3 we made a connection between the length of the intervals and the timing of the tonal peak. We did not make any major statements about the jaw's spatial position because the normalized data did not allow such an analysis. However, the interpretation we made of the interval durations may be slightly rephrased if we apply the knowledge on spatial position that other researchers have developed. The jaw often has a lower position when there is a lower tone (Hirose, 2013; Honda et al., 1999). Since we found that the jaw opening was longer in the A2 condition, this may be an adjustment in position to the initial low tone of A2. Likewise, a higher positioned jaw (=shorter interval) is possible in A1 where the tone initially has a higher tone.

3.4 Paper 4: Mutual influence of word-initial and word-medial consonantal articulation

3.4.1 Summary

Paper 4 is a manuscript that is intended to become a journal article in the future. It differs from the other three papers in two main respects. First, it does not deal with the SWA, although the experiments contain A2 words from the corpus. Second, it is an acoustic study of segment duration and does not include an analysis of the articulation data from the recordings. Still, it complements the other studies by incorporating findings from both Papers 2 and 3: the effect of different word endings on word-initial

CV. Paper 4 can be said to be a more controlled study than the articulation studies since segment duration is a highly established method of measurement. However, the study is still explorative by nature; moreover it is guided by the available corpus material. In addition, Paper 4 contains a detailed discussion on how the results can be linked to articulation. The plan is to follow up the study with articulatory measurements on the same data.

The target words studied were all A2 words with a mixture of word-initial and word-medial consonants, but the vowels were similar, /C1aC2a(r)/, except for the stressed vowels which were either [ɑː] or [a]. Eight target words in the corpus that fit the description were included. The target words were matched in pairs so that only one feature differed between the two words. In this way seven variable sets were created where one consonant (either C1 or C2) was the same in both words (the dependent variable) and the other consonant differed between the two words based on only one parameter (the independent variable). The dependent variables could be either word-initial or word-medial bilabial or dento-alveolar consonants. Thus, the available material allowed us to see if, for example, the place of articulation influenced the subsequent (C2) or the previous (C1) consonant, or both.

Results showed that both C1 and C2 segment duration was significantly different as an effect of the non-adjacent consonant. The effect was most evident when the place of articulation differed between the consonants. Thus, regardless of which articulator produced the consonant, lips or tongue tip, both segments were shortened. This suggests that changing the place of articulation within the word creates a more complex dynamic, which may therefore cause extended duration. Similarly, same place of articulation makes for simpler or more effective articulation, which results in the segments being shorter. Regardless, the results indicate that the position of an articulator, where it travels next, and where it comes from, is significant for the duration of the inherent consonant.

The findings also included inconclusive results on manner of articulation. The influence of manner found on the word-initial consonant may be because of prolonged duration caused by prominence. Therefore, it is not entirely clear whether manner of articulation has a non-local effect on segment duration. Furthermore, the target words that examined manner contained post-syllabic C2, which may have affected these results. A discussion on post-syllabic effects is included in the paper.

A higher prominence seems to also have had an impact on the lack of gemination effect in the word-initial consonant. Previous research has shown that a word-initial consonant is longer if followed by a long word-medial consonant (Turco & Braun, 2016), but we did not find this in Swedish. However, many factors, not only prominence, may have affected the result and we recommend further analyses.

Furthermore, the analysis showed that long consonants and long vowels vary more than short ones, and that word-medial consonants vary more than word-initials.

3.4.2 General discussion

In order to understand the systematic articulatory movements, we must relate them to what the listener hears. This quest begins with the present acoustic study and will be followed by an articulatory study. Paper 4 examines how the connection between different articulatory movements is affected by a phonotactic change within the word.

First, some background to Paper 4. In unpublished analysis work in connection to Paper 2, we found that word-medial consonants have an effect on the CV time lags. There, the word accent pairs /ˈmɑːlɛn/-/ˈmɑːlar/ had greater CV time lags (onset-to-onset) than /ˈmɑːnɛn/-/ˈmɑːnar/. This ultimately meant that in our data C2 had an anticipatory effect on word-initial articulation. This was the reason for initially examining the articulation of the lips and, by extension, the consonant segment.[10]

In what later became an acoustic study, the dividing and comparing of different features of a consonant emerged as a largely successful trait. As Paper 4 shows, some features seem to have a more pronounced effect on segment duration than others. Unfortunately, the gemination comparison was not sufficiently controlled in the study and needs to be reviewed again. Perhaps replicating Turco and Braun's study (2016) is one possible way. Regardless, the results of the place of articulation were unambiguous. It is obviously possible to predict how place of articulation can affect non-adjacent segments. We just need more studies to solve the puzzle.

Prominence seemed to have affected our results after all, as was predicted in Chapter 2. It is clear that despite the attempts to control sentence accent, the target sentences, or the instructions, were not sufficiently clear to the speaker. It is a challenge to do research at syllable and segment level as you are constantly influenced by the slightest change in the articulation. This change need not only be because of prominence level; other factors to consider are the inherent speech rate of the speaker, phrasing, the informativity of the word, etc. In Chapter 4, this topic is discussed further as well as

[10] The findings on the anticipatory effects were first raised as a detour during poster presentations at several conferences. The subject was the focus of the study itself at the LabPhon conference in Lisbon 2018 (Svensson Lundmark & Frid, 2018). In connection with the work there, it was discovered that both articulation and segment duration were influenced by the subsequent consonant. For a while after that, I tried to find a way to explore this issue further. I was fortunate that Professor Martine Grice agreed to serve as my mentor and to help me get off to a good start. She advised me to do an acoustic study, which could then guide an articulatory follow-up study. Professor Grice was deeply involved in designing Paper 4. Unfortunately, due to unforeseen events she could not participate in work on the final phase of the paper.

how we can change our way of working with the influence of different parameters on the segments.

The idea of an articulatory study is still on track. The results of the acoustic study have clarified what needs to be looked at next. For example, we need to see what makes C1 and C2 shorter when the place of articulation is the same. Is it because the movement is faster, or does the travelled distance affect the segments? In addition, it is necessary to ensure that the articulation of the word-initial consonant is indeed different between two words such as /ˈmanːa/ and /ˈmamːa/. If not, we have to reconsider why the segments differ. Work on the projected follow-up study is ongoing. Preliminary results show a complementary relationship between segment length and articulation intervals, i.e. that shorter segments display longer intervals (Svensson Lundmark, 2018; Svensson Lundmark & Frid, 2018). The remaining sections in this thesis will now return to the overall results of all four studies for further discussion.

4 Discussion

What do the results of the studies mean? How do they fit into the larger thinking on these issues? This chapter tries to answer these questions based on some defined thematic boundaries. These themes are the coordination between articulators, an articulation-acoustics relationship, assorted articulatory measurements, and articulatory efficiency in communication.

4.1 Cohesion between articulators

I start with the interaction between the various articulators. This issue has been partially touched on in Papers 1 and 2 as I discussed the measurements of gestural overlap, or time lag. Paper 2 also discusses the results partly from the perspective of interaction between articulators. Beyond that, interaction itself has not been the main focus of any of the papers. However, it is appropriate to discuss the overall results from this point of view.

There is an obvious interaction between all orofacial parts that comes from the fact that they are biomechanically linked. The tongue is connected to the jaw, both of which are connected to the laryngeal structures by the hyoid bone. The lips in turn function as extensions of the jaw; the lower lip belongs to the lower jaw (the mandible) and the upper lip to the upper jaw (the maxilla). This in turn means that even though the lips have different degrees of freedom than the jaw, they are also attached to the laryngeal structures via the jaw. In addition, both the upper jaw and the upper lip are closely connected to the head movements, and act in correlation with, for example, nodding.

The laryngeal structures in turn consist of a complex and well-timed network of muscles, cartilages and joints (for an introduction, see Hirose, 2013). These contribute in various ways to the structures in the language, such as the distinction between vowels and voiceless consonants through muscles that bring the vocal folds together and apart. f_0 changes, on the other hand, occur through either primarily the CT muscle which tenses the vocal folds for a higher pitch, while lower tones require the vocal folds muscles to relax (see e.g. Hirose, 2013). This seems to be done most effectively by lowering the entire larynx (Honda et al., 1999). Such a manoeuvre depends on other extrinsic muscles which in turn are the same that seem to be activated when moving the hyoid bone connected to the jaw (Erickson et al., 1983).

Likewise, the different positions of the tongue in the mouth affect the laryngeal structures. Evidence of this can be seen in the intrinsic pitch of vowels, i.e. in the correlation between f_0 and vowel quality (according to the tongue-pull hypothesis, see Ohala & Eukel, 1987, for an overview). Thus, it is reasonable to assume that, through the biomechanical coupling via e.g. the hyoid bone, change in the laryngeal structures can affect the movement of the tongue, and the jaw, and vice versa.

Thus, all organs in the oral cavity seem to influence each other, either as secondary effects, or through synergetic cooperation, as auxiliary articulators, or as compensation even. Despite this, it is usually assumed that the articulators themselves have a purpose of their own, and that they work relatively independently of one another (Saltzman & Munhall, 1989). Furthermore, it is often assumed that there is a further relationship between them, which is not due to purely biomechanical coupling, but relates to timing between the articulators (Saltzman & Munhall, 1989; Browman & Goldstein, 1992).

Such interaction between articulators is often referred to as e.g. *inter-articulatory cohesion* (Mooshammer et al., 2006), *intergestural cohesion* (Saltzman & Munhall, 1989), *gestural phasing* or *coordination* (Browman & Goldstein, 1992), or *inter-articulatory programming* (Löfqvist, 2003, 2006). The strength of this type of cohesion is usually judged based on how well the different movements coincide, i.e. timing (Mooshammer et al., 2006). Some aspects seem less prone to variability than others and are therefore more or less time-locked with each other. Peak velocity of both upper and lower lip has, for example, been shown to be timed with peak velocity of the jaw during the bilabial closure (Gracco, 1988).

In Paper 2 of this dissertation, we found timing between the acceleration of the lips and the tangential onset of the tongue body movement of the vowel. This indicates that inter-articulatory cohesion is strong between these articulators at a given time. The assumption is that such timing may indicate a phonological function. In the findings we made, it is unclear how well such timing stands against different types of context. For example, we could see that SWA had a small but significant effect on it. Löfqvist and Gracco (1999) found instead that the coordination between lips and tongue varied between different types of consonants and vowels; in addition, there was a certain temporal window (<50 ms), seemingly speaker-dependent, within which the variation occurred. In our material we do not vary the types of CV sequences, as Löfqvist and Gracco (1999) have done in their study with the same measures. Furthermore, we have not analysed speakers individually and therefore do not know whether there are similar temporal windows for how tones affect the timing found. It is thus unclear how well synchronized timing needs to be, in order to be considered as having strong cohesion.

Perhaps that particular measurement in Paper 2 is not sufficient clarification, or timing is indeed variable within a specified time window.

In Papers 1-3 we have only examined bilabial consonants in connection with vowels. In these cases, we assume that the articulators are more or less independent of each other. Thus, the bilabial stops /m/ and /b/ do not have a direct impact on the tongue body movement, or vice versa, regardless of whether the jaw is affected. However, this may differ from person to person; some speakers exhibit more independent control between lips and jaw than others (Kawahara et al., 2014). Nevertheless, the link between the articulators studied in Papers 1-3 is thus mainly about how the jaw correlates with the lips, on the one hand, and with the tongue, on the other. However, we should not forget the inter-articulatory cohesion that is the focus of three of four of the dissertation papers. I am referring to the relationship between the jaw, the tongue and the laryngeal structures.

4.1.1 How does tone affect inter-articulator cohesion?

In Paper 2, a connection is made between tone and the tongue body, in Paper 3 between tone and the jaw. The effect in both cases seems to be two-fold: the position of the tongue may affect f_0 outcome, but an increase of f_0 also entail a higher jaw and thus a probable secondary effect on the tongue trajectory. This dualistic connection becomes clearer if we make a comparison with the relationship between the articulators and the formants. In the case of formants, it is rather a one-way impact; we cannot claim that, for example, a low F_2 is an articulatory target; rather it is the result of a tongue tip constriction that causes the tongue body to move backwards. The fact that f_0 is an acoustic expression of one's own articulator makes cohesion between articulators and structures enabling f_0 more complex.[11] Formants are altered by the inter-articulatory cohesion, on the basis that the articulators themselves are affected, while the laryngeal structures affecting f_0 is both infused and affected by the cohesion. By that is meant that they are not only affected by the mechanical coupling but also presumably timed with the other articulators. In other words, it is not possible to draw unilateral conclusions about an articulator's possible impact on or influence of f_0. Since f_0 is precisely the expression of "its own" articulator, it is in its cohesion with each of the other articulators that we can come close to an answer to the question posed by the title of this section.

So, as far as the tongue is concerned, it has already become relatively clear that the tongue and the laryngeal structures are in a mutual relationship due to the biomechanical coupling. The same seems to apply to the jaw and the laryngeal

[11] Actually, there are different articulatory movements for f_0 raising and f_0 lowering (as mentioned in Chapter 1), but for the sake of argument, let us assume that f_0 is represented by only one articulator.

structures. When it comes to the SWA, that is, the target words of the studies, they consist of several tonal movements on primarily the stressed syllable. During a syllable a lot happens with both jaw and tongue movements: the jaw is lowered and raised, and the tongue moves according to the constrictions as it is supposed to do (as determined by the phonematic structure of the word). The different articulatory movements mean that the cohesion between the articulators also changes over the course of a syllable. These changes in cohesion between articulators also apply to the structures enabling f_0 changes. The results of Papers 2 and 3 indicate that we see greater cohesion between laryngeal structures and the jaw at the beginning of the syllable and a greater cohesion with the tongue at the end or middle of the syllable, that is, in conjunction with mid-vowel or the vocalic target. In the same way, one can suspect that the lips seem to interact with f_0 at the beginning of the syllable because of the connection with the timing already shown between the upper lip, lower lip and jaw velocity (Gracco, 1988). The inter-articulator cohesion between articulators, both those enabling changes in f_0 and others, can therefore be described as a *variable timing*. By variable is meant that the laryngeal structures vary in inter-articulatory cohesion during the course of the syllable and how it is timed with different articulators.

4.1.2 Implications for the Swedish word accents

What, then, does this approach mean for the SWA? Such a description of a variable inter-articulator cohesion with structures enabling f_0 would basically suggest that the initial tone of the SWA may be time-locked with jaw and lip peak velocity, and the second tone with the tongue body. A factor that strengthens this hypothesis has already been found for two speakers of the southern Swedish dialect in Swedish: lip peak velocity was correlated with the initial high tone but only for A1.[12] By extension, this would mean that the tonal patterns of SWA are partly dependent on the movements of the various articulators and on which articulator is used. In other words, a variable timing with articulatory changes of f_0 entails that the f_0 height is adjusted depending on the consonant and vowel type. This has already been documented in several studies on the SWA (Löfqvist, 1975; Öhman, 1965; Elert, 1964).

Furthermore, a still unresolved question is how the f_0 fall and the f_0 rise, which are actually expressions of two different articulatory movements, in turn affect either inter-articulator cohesion. This is undeniably a matter for future research. However, a preliminary analysis can be made based on Öhman's pulse model (1967). The early presence of creaky voice in A1 (Paper 1) shows that the word accent is, so to speak, finished at the end of the stressed syllable. Thus, A1 probably ends with a negative

[12] The study is a conference paper from ICPhS2015, and follows the articulatory development of the pitch-to-segment tradition (Svensson Lundmark et al., 2015). However, the study is not included in the dissertation because of errors in the study regarding measurements at the end of the syllable.

pulse, or a relaxation of vocal chords. Such a manoeuvre is made with a lowering of the larynx and the jaw (Erickson et al., 1983). This should therefore be related to mechanical prolonged intervals of both the lips and the jaw. Unfortunately, we cannot draw such conclusions from the results in Paper 2 and 3: both the lips and the jaw are seen to be affected by the SWA, but we have not specified their timing with the f_0 fall. Nevertheless, such a coupling may be purely mechanical, and not necessarily a case of inter-articulatory cohesion with the articulator lowering f_0. A further complication is the low initial tone in A2, which, according to the pulse model (Öhman, 1967), is an early negative pulse, superimposed on a tonal rise. Initially in A2 there is therefore a simultaneous lowering and raising of f_0. Perhaps this phenomenon explains the articulatory differences found between A1 and A2. This could be developed into another, alternative, explanation, than the one we present in Paper 2. Modelling on how the timing of the laryngeal structures occurs with f_0 could be helpful in order to test this explanation, and to test the hypothesis that we have variable timing depending on the place in the syllable, but also depending on whether it is a low or high tone.

The discussion presented here, and the results, can be further supported by related studies in tonal articulatory alignment. This line of research is about cross-linguistic articulatory work on prosody as well, but rather follows the pitch-to-segment research tradition and asks how f_0 is coupled to articulatory targets instead of acoustic segment boundaries (see e.g. Gao, 2008; Niemann et al., 2011; Mücke et al., 2012; D'Imperio, 2007). For example, German pitch accents show a stable alignment with the vocalic gesture, especially between the tone peak and the target of the vocalic gesture (Niemann et al., 2014; Niemann & Mücke, 2015). Rather, the studies on SWA in this thesis have followed another prominent line of articulatory work, which is generally integrated with basic research on speech motor control and coarticulatory movements, aiming to understand how consonant and vowel articulation is affected by, among other things, prosodic conditions. That both these research directions of tone and articulation seem to lead to similar conclusions suggests strength as regards both theory and results. It also shows that the degree of testability is high in these research areas and that the artefacts (which we do not want) are therefore easier to dispose of. Thus, working with articulatory data in this way is a possible and highly viable way for prosodic research.

4.2 Linking articulation with acoustics I

Another major thematic topic that this thesis studies revolves around, but which is left unaddressed to a large extent, is the relationship between acoustics and articulation. This relationship is also partly the relationship between phonology and phonetics. Phonetic sciences often look for the static and invariable acoustic parameters of dynamic and highly variable articulation. In this dissertation's studies, this meeting takes place only to a small extent. In Paper 1 we found, among other things, a link between lip closure offset and formant changes. In Paper 2 we began to look for systematic differences in the articulation itself. There we were guided by the certainty that timing in inter-articulator cohesion can be phonological. This newly gained insight guided the work in Paper 3 and enabled us to find systematics in jaw acceleration: vowel length is governed by the duration of the open jaw posture. In Paper 4, we again focussed on measuring acoustic data with focus on duration. In the discussion that followed, we addressed the articulatory coordination that the measured segment differences could be due to. That is, we have been looking for systematics but not necessarily with a focus on acoustics. A natural next step is to ask oneself: how can articulatory systematics find acoustic expression?

In Chapter 1, I address how dynamical systems form the basis of our speech. Although dynamical systems govern various physical local elements, such as air vibrations, the four studies are not concerned with these local structures. Rather, they deal with the movements of the larger masses of the articulators. In this section (and in 4.4. *Linking articulation with acoustics II*), I therefore discuss some acoustic connections to the movement patterns of the articulators, i.e. their position, speed, mass and acceleration. One of the available articulatory-acoustic models, the *Locus equation,* may be directly related to our results because of its focus on speech dynamic movements.

4.2.1 Locus equation as an explanatory model for articulation

Locus Equation (LE) is basically about linear regression models. It involves the intercept and slope of a movement, and thus predicts the articulatory movement (Lindblom, 1963). It has been related to formant transitions during the vowel (Lindblom, 1963), but has also been used as an F_2 measure at consonant release correlated to the consonant's place of articulation (for a review, see Harrington, 2013). Although LE appears to be good at distinguishing between different places of articulation, less clear patterns have been shown for manner (Harrington, 2013). It has been argued that it both distinguishes consonants and is linked to their coarticulatory resistance (for an overview, see Iskarous et al., 2010). However, LE is rather directly correlated to the actual movement of the tongue back, and may therefore be a secondary prediction of place of articulation of consonants, and of coarticulatory resistance (Iskarous et al., 2010). While the LE parameter may still be under debate, it is reasonable that a second order equation involving intercept and slope should be able to meet the need for information about the location and velocity of the tongue body.

The following discussion has sprung from the need to apply an already available model of how articulation and acoustics are related. Thus, the applicability of LE has not been tested in this dissertation. However, LE may be an interesting model for the results of the tongue body movement in Paper 2. There we found that the tongue body seems to be more retracted in A2, although we are not in a position to determine whether this is the result of a changed position or changed speed. Either way, the tongue body displayed a different trajectory and what seems to be a different slope. As we propose in Paper 2, the duration of CV time lags might be a secondary effect of the tongue body trajectory and presumably also the tongue body slope. Therefore, there may be a correlation between CV time lags and tongue body shape and thus also LE.

If such a correlation does exist, it may serve as an explanation of CV time lag differences, in our study and others. A similar explanatory model on CV time lags correlation with tongue body has been proposed by Shaw and Chen (2019). They explain the different CV time lags as a result of the position and movement of the tongue; mainly that the more anterior the tongue is, the shorter the time lag (Shaw & Chen, 2019). However, this explanation does not agree well with our data. In Figure 13, this is illustrated by a comparison between two target words by one Swedish speaker: /ˈmɑːnɛn/ and /ˈmanːɛn/. The short [a] is a more anterior vowel than [ɑː]. The front short vowel results in an earlier target and thus a higher speed, as well as a steeper slope (Figure 13). As is evident in Figure 13 the short vowel corresponds to longer CV time lags than those of the long vowel. Thus, a correlation between LE and CV time lags would indicate shorter time lags for less steep slopes. Likewise, a steeper slope (and thus a faster tongue body movement) results in a longer CV time lag.

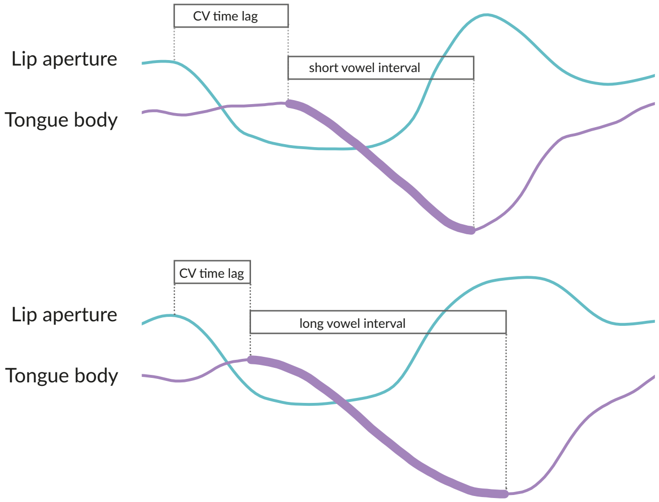


Figure 13. The correlation between CV time lags and tongue body slope.
Trajetories of lip aperture and the TB sensor over time. CV time lags are measured onset-onset between lip closure and tongue movement. The steeper slope of the short vowel gives a longer CV time lag than that of the long vowel.

Judging from Figure 13, moreover, the tongue body movement of the short vowel begins later than the long one. Thus, a putative correlation between LE and CV time lags may have an explanation in the different timing of onsets of the tongue body between a short and long vowel. This explanation also agrees with the word accents; A1 exhibits a shorter tongue body movement (and presumably a steeper slope) as well as a slightly later tongue body onset than A2 (Paper 2). Such a timing difference in the onset may be language-specific and thus not necessarily contradict the correlation between an anterior tongue and CV time lag (Shaw & Chen, 2019).

The tongue trajectory is, however, not the only factor that can affect CV time lag. The point at which the lips close in relation to the start of the tongue is also significant. Moreover, in Paper 2, we found conflicting results regarding the duration of the different ways of measuring time lags. More on time lags can be found in section 4.3.1. For now, we can note that an evaluation of the tongue's role in CV time lags is needed; LE might be a possibility to do so.

Paper 4 suggested that the results on consonant segment duration could be explained by a changed articulation as a result of surrounding non-local segments. A prediction of slope and position may be helpful in determining whether the articulators actually change speed or not. Thus, the movement of the lips either has a changed velocity (affecting slope) and has a different starting position (intercept), or position and velocity both differ as an effect of non-local segments. A more detailed reasoning on this can be found in section 4.4.3. Furthermore, it is possible that LE may be used to predict different trajectories of not only the tongue body but of other articulators as well, such as the jaw and the lips.

4.2.2 Formant changes that lead to a new way of thinking

If we assume that LE can represent the different trajectories of the tongue body between word accents, it seems reasonable to expect that formants also differ between word accents. Thus, if LE can be a possible explanatory model for the tone's effect on the tongue body, we might see lower F_2 in A2, since A2 seems to have a more retracted tongue body. However, F_2 is also influenced by other parameters such as lip rounding, and it is not obvious that this particular phenomenon occurs for all speakers (Harrington, 2013). In addition, lower tones may be predicted for lower positioned tongue through the biomechanical coupling between tongue body and the laryngeal structures. Perhaps a combined explanation model is likely for a phenomenon such as simultaneous low f_0 and low F_2; one that involves calculating both the direction and location of the tongue's movement and its biomechanical impact.

Linking articulation with acoustics is thus not so straightforward that a specific movement gives a certain formant pattern (Jakobson et al., 1969). Furthermore, as far as we know, it is not proven that SWA exhibit different formant patterns. The formants have not been widely analysed in this dissertation. In Paper 1, formant measurement was made on only two speakers. One of the speakers displayed different formants F_2

and F_3 at vowel onset between the SWA. For the other speaker, accent-specific formant transitions were linked to the closure of the lip movement. Thus, formant differences between A1 and A2 existed around the CV phoneme boundary, but they were speaker-specific, at least for our two speakers.

Since we only measured formants at the C-V segment boundary, and not the formant transitions during the vowel, we cannot draw any conclusions about the general formant pattern of the vowel based on Paper 1. However, given the discussion above on inter-articulatory cohesion, perhaps the "frame-by-frame" link between formant and position is actually not significant. Perhaps it is at the segment boundary that the crucial information can be found.

It has long been well known that segment boundaries are characterised by altered acoustic patterns, and thus altered articulation (Fant & Lindblom, 1961; Jakobson et al., 1969; Zsiga, 1994). If we assume that numerous simultaneous articulatory factors affect the acoustics, and we know that the slightest adjustment anywhere changes the inter-articulatory interaction, it may be doomed to acoustically map every part of a segment.[13] Instead, we should adopt a greater perspective and understand the forces that are in the making. This means that we could reflect further on what triggers the actions, and what are the targets. In theory, the targets should have significant functions in the physical movements of the speech, that is, they are either tasks to be performed (Saltzman & Munhall, 1989), or they may function as via-points (for a review, see Perrier, 2012). Articulation is part of the body, and they both follow the same rules and natural laws. We who speak and listen are aware of the conditions of nature, and we act accordingly. As listeners we are thus able to reinterpret the dynamic acoustic outcome based on knowledge of tasks or via-points, but we may still need cues for that. In section 4.4 I develop this reasoning further and present a theory for how to link articulation to acoustics based on time points of rapid formant transitions, that is, at segment boundaries. In order to follow this reasoning more easily, I next provide some background in the form of reviews of some articulatory measurements.

4.3 Evaluation of some selected measurements

Various articulatory measurements have been used in this dissertation. Three in particular have emerged as important, for various reasons, in my understanding of our dynamic and distinctive language; CV time lags, acceleration intervals, and peak velocity.

[13] But, see Iskarous (2017) who argue that measuring only articulatory landmarks can be limiting since local differential equations include discreteness in themselves and can predict movements.

4.3.1 CV time lags

I begin with a consideration of CV time lags, as this measurement is mainly referred to in this dissertation. In Paper 1 it is also referred to as *gestural overlap*, but in the literature one also finds terms such as *gestural timing, gestural alignment, time-locked, timing, temporal coordination, temporal phasing, temporal lag*, or simply *time lag*. Sometimes the timing is specified; e.g. *onset-to-onset lags*. It is sometimes also called *inter-gestural coordination* or even *coarticulation*; if so, clarification of what the measurement consists of is usually offered. It can be confusing that a fairly widespread measurement method has so many names. One reason for this wealth of terms may be that researchers who study the time aspects of articulation do not follow the same phonology. Be that as it may, the timing term hints at how the overlapping articulatory intervals are related to each other over time. In other words, how strong the cohesion is between articulators.

Both Paper 1 and Paper 2 are exclusively about syllable onset, i.e. the word-initial CV sequence. Since, in our material, this constitutes a singleton consonant, in combination with a vocalic gesture, and f_0, there was initially no need for measurements other than just onset-to-onset time lags. In Paper 2, however, the idea was that a more accurate measurement, which replicates earlier studies, would make possible a closer phonological analysis of the articulatory gestures; the bilabial closure, the palatal wide, and the tone gesture. We measured time lags in Paper 2 using zero-crossing velocity; with a threshold from zero velocity to peak velocity (which basically means avoiding the stationary plateau, and instead represents the start of the movement); and also, with acceleration. What we did not measure there, but would have been interesting to see results from, was peak velocity (see more below on peak velocity, 4.3.3.). However, during the work on Paper 2, it soon became apparent that the different ways of measuring timing gave disparate results. This led to the focus being more on understanding how coordination really was than on completing the anticipated phonological analysis. One question that naturally follows is: when does an articulatory gesture begin; is it when the articulators are near still, or when they have started to move? Or is it perhaps when an external force is applied, as in acceleration?

To make a long story short (for a fuller account, see the discussion in Paper 2, and the recap in Chapter 3): it is still a bit unclear what the CV time lags actually measure and what they show. Timing between different features of the articulation can be a significant discovery, but all timing may not be phonologically interesting. It may make sense to measure the lips mid-plateau when the mouth is open, but we should understand why. It may not be possible to overestimate the importance of understanding the timing of the articulation. It is when we know how the articulatory movements relate to each other that we can understand the phonology. What does an articulatory gesture really consist of? When does a gesture start and when does it end? The threshold to peak velocity landmark is almost arbitrary, and it denotes neither one thing nor another. What is competitive between two different gestures? Is it maybe the external force? Since this was still a mystery to me, in Paper 3 I switched rather quickly to following up acceleration and deceleration instead.

4.3.2 Acceleration-based intervals

Löfqvist and Gracco (1999) followed up several previous studies on coarticulation when they examined the spatio-temporal timing between lips and tongue. They chose to use maximal acceleration as a landmark for the onset of the lips as it proved to be most stable between the overlapping various rounded and unrounded vowels.[14] Hence, acceleration was investigated as a timing landmark in Paper 2 due to the study by Löfqvist and Gracco (1999). Since "their" time lag not only presented significant difference between the SWA in our study but was also the least variable of the measurements used, it became natural to investigate it further. In addition, the lip intervals based on acceleration and deceleration were also significantly different between the SWA, indicating a possible phonological significance.

4.3.2.1 Acceleration and deceleration phases

In general, one often encounters the terms *acceleration phase* and *deceleration phase* (not be confused with the term *acceleration interval*, which is often referred to in this thesis). The terms acceleration and deceleration phase refer to departure to and from peak velocity, respectively. The acceleration phase and deceleration phase are different in function and may also be different in shape. Both these phases are based on Fitt's law, which is, in short, a relationship between speed and accuracy (for further discussion, see, for example, Bootsma et al., 2004). There may be an asymmetry in the acceleration phase and deceleration phase due to stiffness and damping, which in turn depends on how difficult the target is to get to and how far away it is (Bootsma et al., 2004). Thus, during the acceleration phase there is not as much need for accuracy; on the contrary, the speed can increase if the distance is long. Instead, when peak velocity is reached, the deceleration phase is instead determined by how large or how difficult the target is (Bootsma et al., 2004). If the target is difficult to reach, the speed then needs to be lower in order to adjust. Hence, the acceleration phase and the deceleration phase need not be equal length. When the acceleration phase is shorter, it also means increased

[14] In Löfqvist and Gracco (1999) the term is referred to as “minimum”: “/…/ the onset of the lip closure for the stop consonant was defined as the minimum in the second derivative of the lip opening signal prior to the oral closure, cf. Fig. 1.” (Löfqvist and Gracco, 1999). This specification refers to a mathematical concept of local minimum, i.e. when the curve goes below the 0-line. However, I have decided not to use the mathematical terms local maximum and local minimum acceleration, but instead *maximal acceleration* and *maximal deceleration* (which can be interpreted more qualitatively) to better visualize their function in speech (similarly, zero acceleration is referred to as *minimal acceleration*). In Paper 3, which deals with jaw movement, local maximum coincides with maximal acceleration. With regard to lips, maximal acceleration coincides instead with local minimum because LA is a calculation and not the actual position of the sensors.

stiffness, while when the deceleration phase is longer, this indicates smaller targets, and damping becomes greater (Bootsma et al., 2004). Task difficulty therefore increases the asymmetry between the two phases, because it increases both stiffness and damping.

As already stated, neither the acceleration phase nor the deceleration phase has been used in this dissertation. These phases are useful when comparing different manners of articulation and may be a possible way forward towards understanding the results in Paper 4, for example. The relationship between these two phases may also be useful when analysing the tongue body data in Paper 2. In Paper 2, it seems that asymmetry is right weight at short vowels; the acceleration phase is longer, and the deceleration phase is shorter (see Figure 14). Long vowels instead seem to have a left-weight asymmetry; shorter acceleration phase and longer deceleration phase. This might demonstrate that task difficulty is greater for long vowels, since it shows both higher stiffness and more damping. Which in turn might mean that long vowels are more articulately complex than short vowels in Swedish. This pattern of acceleration and deceleration phases is similar to that seen between the German vowels tense and lax (Kroos et al., 1997). However, this hypothesis about the Swedish vowels is based on a qualitative assessment and needs to be tested on the data.

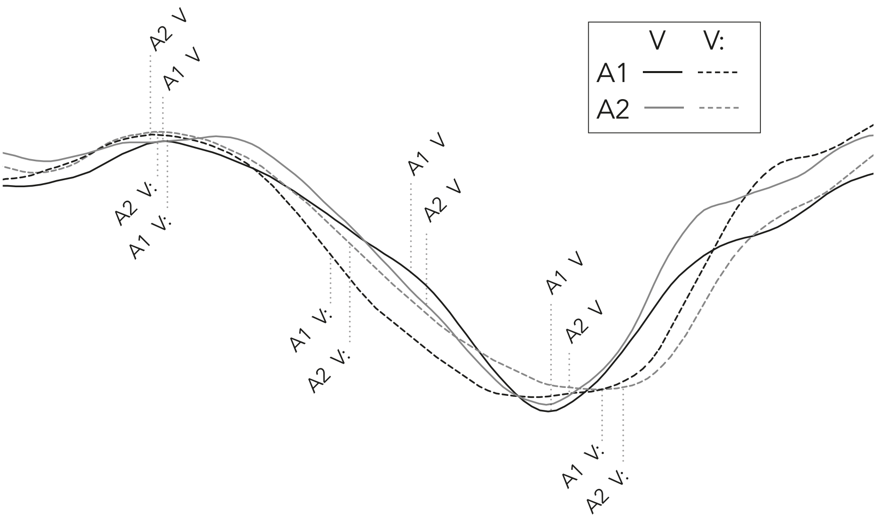


Figure 14. The tongue body height landmarks in Paper 2.
The figure shows where the three tongue height landmarks (based on maximal de-/acceleration of the tangential movement) were approximately vertically positioned. Measuring points and vertical TB trajectories are from four different target words by one speaker. The four measuring points in the middle, which thus show minimal acceleration (= peak velocity), differ the most between the words with short and long vowels, and result in differing lengths of the acceleration phase and the deceleration phase, respectively.

However, there is no agreement in the literature as to what counts as the start of the acceleration phase, and even when deceleration is complete. In theory, zero velocity is the starting point of an acceleration phase. Zero crossing velocity has been used as a measure of the start of closing or opening a gesture in *time-to-peak velocity* measures (which is another name for acceleration phases) (see, e.g., Byrd et al., 2005; Cho, 2002; Mücke & Grice, 2014). However, movement plateaus make the zero-crossing velocity point whimsically placed and therefore may not at all capture gesture onset. Cho (2002) discusses how to manage a plateau with multiple zero-crossings. There the start of the acceleration phase was set at a threshold of 5% (the same threshold as for the end of the deceleration phase). In Kroos et al. (1997), the 20% threshold from zero velocity was used as the onset of the acceleration phase. This was established after experimenting with what appeared to be most stable.

As has been shown in Paper 2 in this dissertation, and elsewhere, the onset of a movement based on velocity is tricky. On the one hand, the 20% velocity threshold is an arbitrary point that seems to replicate when the movement starts to take off. It can easily become a matter of judgment for researchers when they encounter speakers who exhibit motion plateaus that are out of the ordinary, for example, longer postures, which are presumably linked to speech rate. In addition, different articulators of different size naturally operate at different rates of speed and stiffness, which means that any threshold may show different points in different articulators. Although in theory zero velocity is the starting point of an acceleration phase, it only applies under perfect conditions. It is, of course, possible to include the speaker's speech rate in the assessment, as well as other individual traits, but why take into account things that motor control already naturally handles? In other words, there are better landmarks to measure the onset of a movement from. Maximal acceleration may be one.

4.3.2.2 Maximal acceleration and deceleration

Paper 3 makes a radical decision by measuring intervals between maximal acceleration and maximal deceleration. In other words it completely circumvents the function of peak velocity (which in theory is the same as minimal acceleration). This was done as a follow-up to the stable intervals found on both the opening and closing of the lips in Paper 2. It was not entirely obvious that Paper 3 would only use acceleration-based intervals, but it was born out of studying the vertical and horizontal velocity profiles and the acceleration profiles.

But what does an acceleration interval actually consist of, as we have measured it? There are different types of acceleration intervals. On the one hand, we have intervals involving plateaus of no or very little motion, and they should therefore be referred to as *postures*. These may be, for example, the closed part of a constriction, or the open part of the vowel. These (quieter) postures are intervals between maximal deceleration and maximal acceleration. The more active parts of the intervals are those that are instead measured between maximal acceleration and maximal deceleration, i.e. a reverse order of start and end points (see Figure 15). The *active* acceleration intervals are not part of the quieter movement plateau. Instead, they include peak velocity, which is positioned approximately in the middle of the intervals.

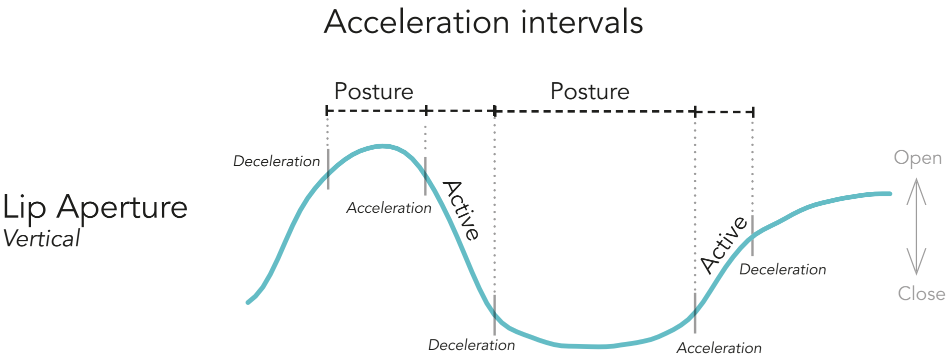


Figure 15. The acceleration intervals.
Specification of acceleration intervals on lip aperture; posture intervals and, between them, active intervals. Landmarks signify maximal acceleration and maximal deceleration, respectively.

In other words, these active acceleration intervals (which, noticeably, are nothing like a quiet posture) consist almost exclusively of an acceleration phase and a deceleration phase (presented in 4.3.2.1.). However, as has already been stated, it is not entirely obvious how an acceleration phase is measured. It follows that it is not possible to confirm that an active acceleration interval consists of exactly one acceleration phase and one deceleration phase. In addition, we know from the research on motor control (see reasoning in Bootsma et al., 2004) that the acceleration and deceleration phases are not always discrete movements; but that the phases can overlap. In fact, by dividing the movement of an articulator into a passive posture and an active part, we may be able to escape some of the aforementioned overlap, which may be problematic in, for example, phonological categorization. Thus, we may also be talking about two parts of the active acceleration interval; the first, which accelerates to peak velocity, and the second, which decelerates after peak velocity. Whether the first can denote stiffness and the second damping remains to be seen.

Through Paper 3, it became clear that it is in the posture of the open jaw that the distinction between the long and the short vowels becomes evident. Furthermore, in both the active part of the opening of the lips and the jaw we found differences between tones. What other phonological distinctions can we find in these intervals? Further research can undoubtedly find out the answer to this question, in order to ask new ones. Section 4.4. develops this idea of acceleration intervals, based on the boundaries between them, that is, in the vicinity of maximal acceleration and maximal deceleration.

4.3.3 Peak velocity

Peak velocity of a movement has previously been used extensively in articulatory research (Gracco, 1988; Cho, 2002; Byrd et al., 2005; Löfqvist, 2005; Brunner et al., 2011; Erickson et al., 2014; Mücke & Grice, 2014; Türk et al., 2017). While I have been working on this dissertation, I have returned to it at various times, but it has not been included as a measurement method in any of the studies that now form part of the thesis, except as an exploratory point for measuring tongue body height (Paper 2). Since it is nevertheless an accepted measurement method, I see it as an opportunity to explore it further in the following paragraphs.

Since peak velocity unite the movements of the upper lip, lower lip, and jaw (Gracco, 1988), one can assume that the highest speed in an articulation has high inter-articulatory cohesion, and as such is a possible phonological function. Perhaps the onset-onset timing of gestures is rather an effect of this upcoming timing in peak velocity? Either way, it is not entirely clear how variable peak velocity is compared to onset-onset timing. What we do know is that onset-onset can vary by approximately 50 ms in speakers (Löfqvist & Gracco, 1999). Even though we did not compare the individual speakers, or varied the CV sequences, we found time lag variations as well (see Paper 2). However, the lip closures based on acceleration varied less than those measured by velocity. It may be that onset-onset timing has a phonological function for one context (i.e. a specified CV sequence), while timing in peak velocity is a function for another (i.e. any CV sequence). On the other hand, it may also be that one or both of them are actually the result of biomechanical connections between articulators. This may specifically be the case for the peak velocity link between the lower lip and the jaw.

As previously mentioned in section 4.3.2.1, task difficulty increases the asymmetry between the acceleration and the deceleration phases, because it increases both stiffness and damping (Bootsma et al., 2004). Sometimes the acceleration phase is also referred to as *time-to-peak velocity*, a measurement method used in many studies since it is a measure of the stiffness of an articulator (Byrd et al., 2005; Cho, 2002; Mücke & Grice, 2014). Thus, the advantage of articulatory measurements of time-to-peak velocity is that it measures not only a mechanically obtained distance, but also a dynamic property of the articulatory phenomenon. Furthermore, the peak velocity landmark is the same landmark as minimal acceleration, meaning when the movement has the highest speed and no external force is needed. It is also after this point that velocity begins to decrease. Since, in a damped mass-spring system, velocity decreases when the target is overshot, peak velocity is also a possible target. Thus, peak velocity has great potential to be a phonological function in speech.

Such a possible phonological function will be briefly mentioned here. An explanatory model of how speech is organized which has not been mentioned so far is the C/D model (Fujimura, 2000). The C/D model acts as a bridge between articulation and acoustics and, like other gestural phonologies, it assumes an organization based on articulation rather than on phoneme segments. In short, it can be said that the syllable is considered to be the supporting organizational construction. Furthermore, the phonological model can account for syllables of different magnitude, i.e. strong/weak or stress level (Fujimura, 2000). Thus, it can be applied to prosodic levels. How this can work is explained by the fact that the base in the construction consists of triangle-shaped syllable pulses, where the center is the nucleus (Fujimura, 2000). The syllable pulses act as a metrical structure that can capture prosodic variation (Erickson et al., 2014).

Since it is considered to be articulatory organized, the structure should also be measurable, as a study by Erickson et al. (2014) has demonstrated. In the study, the syllable pulse was equated with the openness of the jaw (jaw displacement). The triangular base was further calculated based on the consonants before and after the syllable pulse, i.e. surrounded the vowel, which depending on the consonant meant different articulators (Erickson et al., 2014). Thus, peak velocity was measured on the consonant in the syllable onset after its constriction was done (local minimum velocity) whereas, in the case of the consonant in the syllable coda position, the peak velocity of the articulator was measured before its constriction (local maximum velocity). These two measuring points were thus used to calculate syllable base triangles in order to obtain articulatory syllable duration, information that seems to agree well with stress levels and prosodic boundaries (Erickson et al., 2014).

The study by Erickson et al. (2014) points out the many possibilities that exist in articulatory measurements. Prosodic phenomena and metric structures are expressed through articulation and thus it cannot be doubted that their systematics can also be found. Thus, peak velocity seems to be, as mentioned earlier, a good candidate for a phonological time-point. In Paper 2, we measured tongue body height at minimal acceleration (same as peak velocity). This was an exploratory measurement, a result we have probably not finished analyzing. But given the phonological capacity of this landmark, we should take a closer look at these issues; it seems, for example, that long and short vowels differ in when in time the peak velocity is reached (see also section 4.3.2.1). Thus, a possible difference in vowel quantity might be explained by timing of the peak velocity, basically the speed of the tongue.

4.4 Linking articulation with acoustics II

This thesis provides no modelling of dynamical systems but rather aims to show how applicable dynamical systems are for the development of phonological modelling. Perhaps this becomes even clearer when the connection to acoustics is exemplified. The following section presents a link between articulation and acoustics, which constitutes a specific event in time. But in order to arrive at it, we first need to review the difference between timing and duration.

4.4.1 Articulation in time

Different phonological structures appear to find different physiological expressions. Constrictions, for example, depend on the location of the articulator but also on the way in which the articulation is done. This fact will probably not be disputed by anyone. Since we know that the duration of the deceleration phase is dependent on the complexity of the target, the phonotactic placement of a consonant – the syllabic position, that is – determines not only how well or carefully the movement is performed, but also its duration; the more complex an interval, the longer it is. As a consequence of this, the duration of a movement affects the acoustic pattern. It follows that in order to understand the link between articulation and acoustics, we should perhaps pay more attention to what determines when an action is taken than to the "frame-by-frame" acoustic pattern.

That is exactly why inter-articulatory cohesion is significant - because the structure of the timing of articulation describes which actions are taken. And, whether they are likely to be coordinated, that is, simultaneous, or are performed in succession, i.e., sequentially. However, the fact that the articulatory actions are overlapping makes their description particularly problematic: they are potentially related to each other constantly. Thinking of speech as consisting of segments, as beads on a string, may seem much simpler: segments could be conceived of as isolated sequences that follow one another.[15] However, research shows again and again that segments are coarticulated, and that articulation is constantly overlapping (Öhman, 1966; Löfqvist & Gracco, 1999; Recasens, 2002; Shaw & Chen, 2019). In other words, a constriction's movements are highly dependent on the underlying vowel. Therefore, the function of time is not the same between a segmental phonology and a gestural phonology. In a segmental phonology, the focus is on duration, while in a gestural phonology, both timing and duration are of the essence. In fact, the concept of time in a gestural phonology can be said to have a phonological function.

[15] However, it depends on your view of segments; distinctive features are characterized, for example, by being both superimposed and concurrent (Jakobson et al., 1969).

If time in its turn has a phonological function, it is necessary to specify what type of time. For example, timing and duration are certainly not one and the same. Timing describes when something is performed, and it is presumably a significant determination of a phonological structural unit. Whereas duration is the main determinant of the resulting acoustic pattern, due to the role of duration in articulation because of the aspects that determine it, such as positions. In this dissertation, such a division could be made extremely clear by using the measurements as examples: thus, timing is exemplified by the measured CV time lags, or gestural overlap; while the different intervals based on position, velocity and acceleration refer to duration.

Furthermore, timing and duration are dependent on each other. Duration is the results of timing, and if a timing is to occur, the articulators need to adjust their movements to get there. Simultaneously, duration, as a mechanical interval, depends on mechanical properties such as original position, distance to target, obstacles in the way, and dynamic features (humidity, heart rate, elasticity etc). All of this is taken into account and speed is adjusted to reach the goal, which is reflected in the time it takes to perform the action.

Thus, when speaking, we are constantly shifting tempo in order to reach timing. We change the tempo to reach a moment that is significant, a target. In other words, timing is similar to a target. Therefore, understanding what targets and timing are is an essential knowledge of the function and form of language. However, unfortunately, this dissertation has not come up with an explanation of what the targets are. In the following paragraph I will present a way to link the function of time in articulation to time in acoustics. Maybe it can help us, later on, to further understand timing and what the targets are.

4.4.2 Towards a theory of time and segment boundary

Segment duration thus uses a somewhat different time scale than the articulatory intervals. Segments are sequential, instead of simultaneous like overlapping articulatory gestures. Measurements of segment duration are thus also sequential and move from one segment boundary to the next. The fact that the segments follow each other makes them discrete by nature, and also makes them follow each other in time; this enables them to be perceived as phonological units. However, as already clarified, duration is not completely static but varies systematically based on given parameters. One such is the place of articulation in the consonant before or after, according to the results in Paper 4. Since mass is the same in the affected examples in Paper 4, and is based on the sum of position, velocity and acceleration being assumed to be the same, acceleration or velocity seem to have been what has changed so that the segments are affected. The boundaries of a segment should therefore relate to a specific point in velocity or acceleration, or both, that is different depending on an adjustment to the position that precedes, or the position that follows. Exactly such a similarity seems to exist between the segment boundary of the consonant and the boundaries of the acceleration intervals. The figure 16 below shows how the different intervals relate to the segment duration of the bilabial word-initial /m/. The posture interval that occurs between maximal deceleration and maximal acceleration, that is, the posture acceleration interval, corresponds to the C1 segment duration.

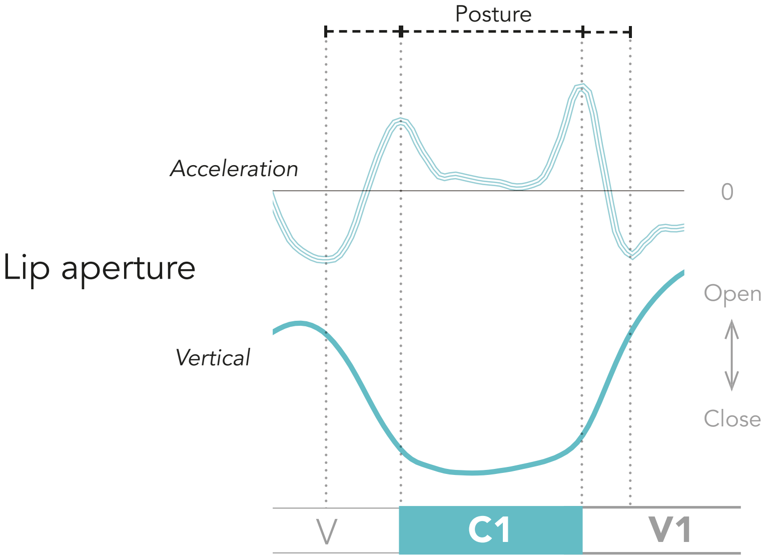


Figure 16. The relationship between posture acceleration interval and segment duration
The figure builds on Figure 15 through the addition of the acceleration profile and the phonemic target. Onset of C1 corresponds in time with lip aperture maximal deceleration; offset with maximal acceleration.

Just as the sequential order of the segments is a phonological strength in a segmental phonology, the sequential order of the proposed acceleration intervals may be a phonological strength. Since they are based on a specific time point (as opposed to the acceleration phase and the deceleration phase that can overlap), they are distinct and static in relation to each other. At the same time, they are capable of overlapping intervals of, and timing with, other articulators. This is similar to articulatory gestures in a gestural score within AP (Browman & Goldstein, 1992). In addition, the acceleration intervals are adaptable to the constantly changing positions and velocity of the other articulators, as they are based on the acceleration of the inherent body. Thus, these intervals are both overlapping and sequential, both dynamic and static. In other words, what appears to be a time-locked relationship between maximal de-/acceleration and acoustics can explain how the function of time in articulation is related to segment duration.

In Figure 16, it is lip aperture that exhibits this phenomenon, which perhaps only applies to the relationship between two independent articulators. However, not only the bilabial segments display this tendency. The tongue tip also appears to exhibit the same relationship between the acceleration interval and the segment. A trained eye can see from Figure 9 in section 2.3.1 that the tongue tip rapidly changes position, and thus probably contributes to a rapid change in velocity (a maximal acceleration or deceleration) at the same time as the segment boundaries. This particular phenomenon is not presented as a result in any of the thesis studies though it has emerged as a recurring pattern at the analysis stage of my work. However, this correspondence between the posture of the acceleration intervals and segment durations has only been present in voiced consonants, since it is the limitation of this dissertation material. The fact that it has only emerged as a feature of consonants may have to do with the overlapping nature of consonants. That is, consonants act like a layer on top of the vowels (Öhman, 1966). However, the consonants are not independent and are influenced by the movement of other articulators due to inter-articulatory cohesion.

This proposed *time-locked acceleration-to-acoustic-boundary* coincides with formant transitions. It is a well-established fact that segment (or phoneme) boundaries occur where formant transitions occur, that is, where speech events are rapidly changing (Fant & Lindblom, 1961; Gårding, 1967; Jakobson et al., 1969; Zsiga, 1994). Thus, maximal acceleration and deceleration not only account for one of the major kinematic changes, but also for one of the major acoustical changes, if not the main one. The fact that rapid changes in articulation have a bearing on the articulatory-acoustic relationship has also been suggested recently by Goldstein (2019). The present dissertation thus suggests that this specific activity, the acceleration and deceleration of a movement, is a highly significant feature of human speech. Therefore, the correlation between the acceleration intervals, and specifically the points in time of maximal acceleration and maximal deceleration, deserves more focus in speech research, not least if it is considered from a perceptual point of view.

4.4.3 A time-locked acceleration-to-acoustic-boundary hypothesis of place of articulation

What does this hypothesis of a time-locked link between acceleration and acoustic boundary mean for the results of different segment lengths in Paper 4? We use the example of the mutual influence of the place of articulation as the tendency at hand is most noticeable here. If segment boundaries are equated with maximal acceleration and deceleration, then either one or both of these points change over time depending on parameters belonging to another consonant's position. To rephrase: if C1 and C2 have the same place of articulation, the posture intervals are shortened. So, with regard to the mutual influence of the two target segments C1 and C2, it is reasonable to assume that it is mainly the active interval after C1 and the active interval before C2, respectively, that is affected. This means, to be more specific, the points of maximal acceleration at C1 offset and maximal deceleration at C2 onset (see Figure 17).

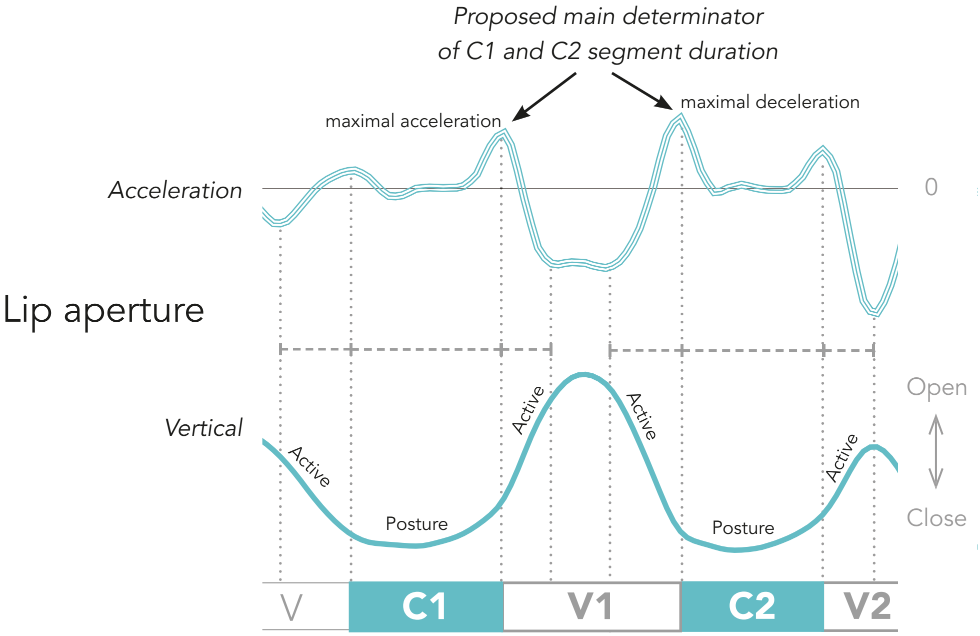


Figure 17. The acceleration-to-acoustic-boundary hypothesis of place of articulation
The two active acceleration intervals of the lips, simultaneous with the vowel (V1), are proposed to determine how long the C1 and C2 segments will be. More specifically, segment length is believed to depend mainly on the onset (maximal acceleration) of the active interval that shows how far the articulator needs to travel, and offset (maximal deceleration) that shows how much the articulator slows down.

Moreover, it seems reasonable that the two time points act independently of each other as they are on "opposite" sides of the posture of the open vowel and thus are acting independently of each other in different active intervals (Figure 17). Maximal acceleration is part of the acceleration phase (or time-to-peak velocity interval), which may have different lengths depending on how far away the next target is, that is, theoretically depends on how fast the articulator needs to be to get there. Similarly, maximal deceleration is part of a deceleration phase that can be longer or shorter depending on the severity of the target.

One might therefore assume that maximal acceleration would start earlier, i.e. shorten segment C1, if there is a longer way to travel to the target so that it needs to move faster (more stiffness). Furthermore, one might also assume that maximal deceleration occurs later, i.e. shortens segment C2, because the target is more complicated (more damping). However, this is not directly in line with our results.

In Paper 4, it is shown that when the place of articulation is the same for both consonants the segments become shorter. In other words, articulation is assumed to be simpler, or at least less complex. Thus, one cannot completely base the effect of the segment shortening on the complexity of the target, that is, the deceleration phase before C2 onset. Still, the fact that the segment is shortening may possibly have to do with the acceleration phase at C1 offset. Since this particular articulator will soon be used again, it must act fast to be able to participate in the open vowel in between C1 and C2. Thus, the same place of articulation possibly allows the articulator to travel longer, i.e. move faster so that the articulator displays more stiffness. Similarly, the acceleration phase in the subsequent active interval towards C2 should be equally affected, i.e. it must travel back the same distance, and be equally fast. It is therefore reasonable to assume that in a CVC sequence the two active acceleration phases of an articulator, those that appear in close proximity to the vowel, are what mainly affects the length of the surrounding consonant segments (Figure 17).

However, another possible explanation of the shorter word-medial consonant segments may also be that the faster movement of the first consonant somehow lingers on and coincides with the second consonant. This is the same explanatory model as the hypothesis of the expected but not found gemination effect in Paper 4. If this were the case, speed might bind a unit such as the syllable together. That is, speed affects nearby and non-adjacent segments, because they are part of one and the same planning unit. However, it is also highly possible that the same functions determine both word-medial and word-initial consonants, simply because if the articulation is not very complex, the articulator can be extra fast. Which of these solutions is correct, if any, remains to be seen.

4.5 Communicative efficiency

The following passage is about articulatory efficiency and contrast richness. It has not been an explicit aim of the individual studies to pursue these issues. Rather, it is an overarching purpose of the whole thesis and has motivated the direction into which my research has moved. Aspects of communicative efficiency were not taken into account when the material in this dissertation was designed, apart from the fact that I knew that I would be studying a controlled set of phonemes, and would focus on the word-initial segments. Still, the result based on this knowledge can be analysed and maybe serve to build a hypothesis.

4.5.1 An “immune” and contrastive articulation

We have already found that word-medial consonants affect the word-initial segments. This can be interpreted as signifying a link between the consonants within a word, no matter what this link is due to. Paper 4 also introduces a discussion of the articulatory influence in both directions, which in the previous section 4.4.3 was partly explained by the hypothesis of the time-locked acceleration-to-acoustic-boundary. However, the introduction (Chapter 1) mentions some significant differences between word-initial and word-medial segments, for example that, word-initially, perceptual contrast is higher. How can these results of Paper 4 work in conjunction with theories about the strong word-initial contrasts? Shouldn't the word-initial segments be immune to influence? As the following paragraph aims to show, it may be that some articulatory aspects are "immune" while others are not.

The jaw opening degree is strongly correlated with the consonant type (Lindblom, 1983; Mooshammer et al., 2007; Kawahara et al., 2014). Spontaneously one might draw the conclusion that it is the jaw opening that causes the bilabial C1 segments to change, since the lips and the jaw is highly correlated. However, in Paper 3, one could see that the type of word-medial consonant does not affect the jaw opening interval itself. It is as if the word-initial jaw opening was only affected by the inter-articulatory cohesion at that time (which includes the relationship with the lips, the tongue, and the presumed laryngeal structures) and not by what happens at the end of the syllable. If we dissect these results further, they do not really mean that the position of the jaw is not different - the position may be different, and also the jaw velocity. The results of Paper 3 really only show that the jaw acceleration interval does not differ depending on the types of word-medial segments that follow. However, Paper 4 clearly shows a link between segment length and type of following consonant. How are these results related?

First and foremost, the length of a segment has a contrastive function for the listener. Different duration therefore has a significance for predicting the context. The listener can predict from the shorter bilabial segment that the following segments in the word are probably also bilabial, since the acceleration profile is different, causing the segments to differ in length. In this way, variation of word-initial segments can serve as a signal to the listener: some words are “easy”, meaning basically short segments, and some words are “hard”, i.e. longer segments. "Easy" words mean both easy to hear, but also to produce, which enables shorter and faster movements.

However, the variation of a word-initial segment is a secondary effect of the articulator having different starting positions or target positions to begin with, which means that the articulator's speed is adjusted to where it comes from or where it is going, just as has been argued in the previous sections. The variation is large, but at the same time there is the systematics, which was shown in Article 3 where word-initial jaw opening was not affected by the word-medial segments. Thus, in the specific case, acceleration and deceleration are constant. They appear to be influenced only by the functions that are currently in place. In this way, it may be possible that acceleration and deceleration is an expression of a phonological distinctive difference in articulation. That is, segment differences may not be phonological distinctions in their overall duration, but it is what determines the segment duration that is distinctive: the acceleration profile. In Chapter 5, on future research, I raise the question of an articulatory hierarchy based on this hypothesis.

4.5.2 On hard and easy words

Segment duration is affected by how simple or difficult an articulation is, and thus how "easy" or "hard" a word is. Easy and hard words are also related to word frequency, perceptual contrast, and articulatory effort (see Introduction, section 1.2.1). However, our target words consist mostly of nasal consonants. Nasals are generally not considered to be strong in terms of perceptual contrast, which is why it becomes difficult to advance the argument of this section. More studies are needed where the results are compared to greater articulatory effort - will the jaw opening interval continue to be independent even when influenced by, for example, word-medial fricatives?

Unfortunately, our results are not conducive to a deeper analysis of informativity. We have only one word with a high lexical frequency; /ˈmamːa/ (*mom*). Not surprisingly, it is also shown to consist of shorter word-initial segment than /ˈmanːa/ in the analysis in Paper 4. But the word /ˈnanːa/ is also shorter than /ˈmanːa/, and these two have the same word frequency in Swedish. Therefore, even though /ˈmamːa/ has a higher lexical frequency, the same place of articulation may still be the reason for shorter segments. Why /ˈnanːa/ is not more frequent in Swedish is perhaps an even bigger issue, since it seems to be an equally non-complex, or "easy", word as /ˈmamːa/.

However, discussing hard and easy words can easily become a circular argument: the segments are shorter because the word has a high frequency, and the word is frequent because it is easier to say (in addition, the labial /m/ is more contrastive than dento-alveolar /n/, hence easier to distinguish). And to close the circular argument: when the word is easier to say, as Paper 4 shows, the segments are shorter. It follows from this reasoning that we should be alert to normalization when making segmental comparisons. Instead, we should take into account the reason why some segments are shorter when planning studies. Thus, the shorter segments are due to significant movement patterns that are necessary to consider. If we normalize the duration, we lose the nuances that can lead us to what the movement goals actually are. Why does the articulator move faster? Well, because the goal is further away or because it has to take a detour. Regardless, there is an articulatory phonological explanation for the shorter segments. Thus, hard and easy words should not be separated, or normalized, they should only be handled with care in analysing, based on their respective conditions.

4.6 Summary conclusions

This chapter has been about deepening the discussions about the results of the individual studies, within some defined subject areas. For example, the discussion of tones' impact on articulation has been addressed. I highlight a possible hypothesis of articulatory timing in the Swedish word accents: a variable timing with the laryngeal structures, dependent on syllable position but also on tonal rise and tonal fall. This is presumably relevant for other Swedish dialects as well (see section 5.2).

I have also mentioned possible explanatory models for how the articulatory results might be linked to acoustic data. In part two of the divided sections Linking articulation with acoustics I and II, a hypothesis was presented based on the fact that the segment boundaries where the most obvious formant changes occur correlate with the acceleration and deceleration of consonants.

Furthermore, chapter 4 has included an analysis of the methodology used, specifically where various articulatory measurements are concerned. This led to the presentation of the idea of *acceleration intervals*, which includes both postures and more active parts of a movement. Chapter 4 also includes discussions about communicative efficiency: one conclusion about hard and easy words has been that they should not be normalised or separated. Instead, they can increase our understanding of how speech dynamics is related to phonological distinctiveness.

The research work during these years has been mostly investigative and this has led to many loose threads that can be followed up in the future. In the next chapter I present some possible ways to proceed.

5 Future research

Work on this dissertation has partly been a fishing trip; I have cast my nets a little here and there and some of my catches have been quite unexpected. Research in the form of a fishing trip is usually seen as something negative in view of its uncontrollable conditions. This may be true, and my goal in the future is to turn to testing the hypotheses presented in this dissertation. However, it seems counterproductive to dismiss the benefits of the fishing trips. Without these, the hypotheses could not have been developed, and without them I would not have been able to find what really seemed most significant. Going forward, I hope to be able to combine hypothesis testing with an occasional much needed fishing trip for purposes of recovery and discovery.

The hypotheses developed in this dissertation are based on a genuine interest in how the body moves and on a lifelong interest in structures. In other words, getting a grip on a well-described model such as AP, which promises to solve the connection between the dynamic and the systematic, becomes an almost personal challenge. What I have not yet managed to understand – I am still busy trying to do so – is how a gesture starts and ends. I encountered this problem in Paper 1, and while I was working on Paper 2 it became clear that the question is too big to answer within the boundaries of a dissertation. My research area seems to have sprung from this specific problem. Thus, understanding how a gesture starts also describes what its force is, which in turn can suggest what its goal is. Grasping the nature of targets is closely linked to grasping when actions are initiated, since speed is determined by what is to be performed.

Perhaps, we should also admit that we may not even know what phonological units are. They may be articulatory gestures; however, if time is not an external function but inherent in the systems, what does it tell us about what phonological units are or how they are executed? Perhaps the question of the function of time will be the big divide in the future. Be that as it may, it is becoming increasingly clear that we cannot manage to develop phonological models without a better understanding of the dynamic systems that control the selection, timing and execution of the smallest building blocks of speech. Continuing to map the mechanical movements can only make a positive contribution to this development.

The following sections introduce potential ways to move on from here. There are many possible avenues for future research, which may be due to the fact that such work places itself in the intersection between many already established research areas. Which, in turn, makes this dissertation perhaps more aligned with interconnected research activities than with in-depth ones. Nevertheless, some important steps need to be taken. The following section aims to highlight a couple of such examples and offer some suggestions concerning unresolved issues in the dissertation. First, I want to discuss how we might develop a framework for the function of time in articulation, including a suggestion of a possible articulatory hierarchy. This is followed by a discussion of how SWA can be further investigated based on ideas about the gesture onset. The chapter ends with an overview of established research areas.

5.1 The significance of articulation in time for modelling speech production

How time is manifested in speech is perhaps the most basic of all the functions of speech that help us understand each other in spoken communication. On the one hand, we have an indispensable tool for conveying sound to the listener over time. Time is not only essential as a revealing function, that is, when something becomes apparent: it is also used structurally when we build up an understanding of what is being said. We utilize this structure because units can refer to each other in time; tones in SWA are, for example, linked to the following suffix. In addition, speech sounds are created by movements overlapping each other in time; that is, time almost acquires the quality of a spatial concept, made up of simultaneous layers of time, each of which can have different meanings for the listener. For example, a long and open vowel, produced with a relatively slow articulator, moves at the same time as a fast lip movement, which in turn can be fast or "faster" depending on its target. Thus, two symbols with different functions figure simultaneously and each according to its "time schedule".

In addition, when a sound is made along the time line, i.e. timing, invaluable information is provided. By the same token a sound can come out completely different if the timing is affected by the slightest of errors (e.g., if the tongue tip does not reach its intended target). Timing can also refer to how different parts of our body can be more or less in timing with each other – for example: is the jaw lowered at the right time or not? And what happens to the quality of the vowel if it is not? As for the individual articulator, for example, the tip of the tongue, the resulting sound will be different depending on how long the articulator travels, or when its peak velocity is reached, which is often referred to as an articulatory interval. An interval is not the same as segment duration, which is rather a reflection, or a result, of the many different approaches to time that the articulation involves. But duration is of course time, too. No one doubts that there is a lexical difference between short and long sounds; in other words, duration is also a significant aspect of time.

Time, thus, seems to perform many different functions in speech. A future functioning phonology should in other words be able to connect the many different functions that time has in speech. Regardless of which aspect of time we are talking about, it would seem that the question of the role of time in speech has not yet been determined. Different phonologies and models assume different things: some believe in an external clock function; others think that time has an intrinsic role in the system; yet a third one relies on phonological specification in representing time (for discussions of this, see e.g. Saltzman & Munhall, 1989; Turk & Shattuck-Hufnagel, 2020; Perrier, 2012). Nevertheless, everyone seems to agree that the function of time is an extremely important – and difficult – issue to solve.

5.1.1 Towards an articulatory hierarchy

One of the many possible ways to proceed with this matter is to look into the connection between the acoustic boundaries and the time-based articulatory landmarks, i.e. the aforementioned hypothesis of a time-locked acceleration-to-acoustic-boundary. For example, we could take a closer look at what this coordination is based on, but also examine how well it might agree with different types of constrictions as well as with other units in the language.

One aspect of this will be addressed here, which leads on to a hierarchical structure. Figure 18 shows a putative synchronization between the boundaries of the acceleration intervals of the lips and the jaw. The figure shows the plotted acceleration curves and how they correlate with the actual trajectory of the sensors (or, in the case of the lips, actually the calculation of lip aperture). Only the three intervals of the lips that are significant for segment C1's duration have been included here, with the posture in the middle. A similar image has been shown in section 4.4.2 (Figure 16). The posture of the lip constriction clearly aligns itself with the boundaries of the C1 segment.

With regard to the jaw, five intervals have been marked: starting with a posture and ending with a posture. In Paper 3, only the three in the middle were calculated (one active + one posture + one active), but two more are included in Figure 18 to display the entire syllable. It is evident from the figure that none of the acceleration intervals of the jaw is timed with the segment boundaries. This is not surprising since the jaw is partly an auxiliary articulator to the constrictions, and also the one articulator that can be said to constitute the syllable.

Figure 18 is indicative of the fact that in the articulatory dynamics there seems to be a hierarchy, as is also expressed in a traditional segmental hierarchy (e.g. syllable > foot > segment). However, for an articulatory hierarchy, this means that articulators that constitute constrictions are at the top, while the jaw do the underlying work for the rhythmic grouping. This proposed hierarchical structure agrees quite well with previous studies on coarticulation (Öhman, 1966) and on the jaw opening function (Erickson, 1998). The overlapping position of consonants seems to be the reason for the visible (and audible) acoustic boundaries.

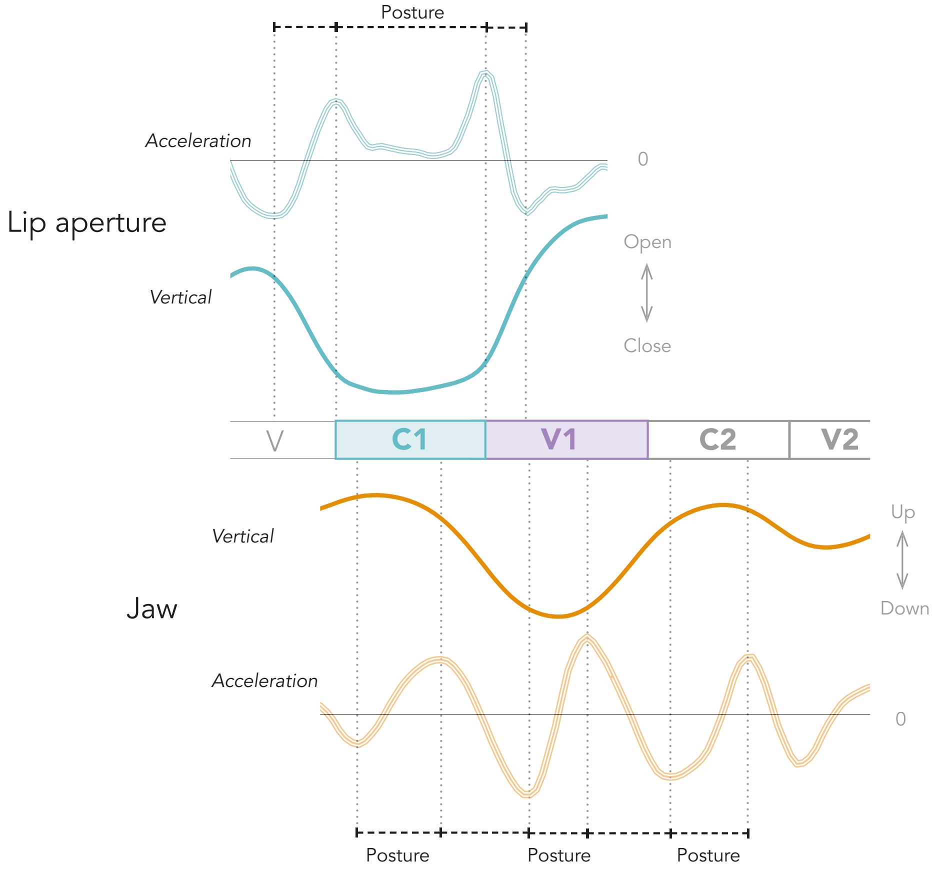


Figure 18. An articulatory hierarchy
The proposal of an articulatory hierarchy is based on the overlapping of gestures. The consonantal gestures create the boundaries of the segment and are thus at the top. The slower jaw has a cohesive role and is at the bottom of an articulatory hierarchy.

The tongue body is not included in Figure 18. This is because no hypothesis has been developed about the tongue, as acceleration of the tongue body is not investigated in the dissertation. This is in turn, partly, because the tongue body is physically complex with many degrees of freedom and thus has the ability to change its speed to a greater degree than a mass with fewer degrees of freedom (such as the jaw). Methodologically, it is more difficult to work with an analysis of the acceleration curve, because of the fluctuating signal, than it is, for example, to analyse the trajectory of the tongue body sensor. However, due to the complex movement system of the tongue, it may be less appropriate to filter the signal so that acceleration can be easily measured. An overly filtered and smoothed signal can increase the estimates of data distortion and thus result in inaccuracies (Hoole & Zierdt, 2010; Hoole et al., 2014).

However, because the tongue body has many degrees of freedom it should be able to be functional in speech in many different ways. When we are able to distinguish between the different functions of the tongue body movements, which will be the case when we have a better understanding of them, we should be able to more easily analyse the acceleration intervals of the tongue body.

Nevertheless, theoretically, the tongue body might figure in a plausible hierarchy somewhere between the fastest articulators (lips and tongue tip) and the jaw (Figure 18). More research on the forces of the movements, such as the acceleration and deceleration of different articulators, is needed to determine whether this hypothesis of an articulatory hierarchy is correct.

5.1.2 Explaining acoustic segment phenomena

Another way to increase our understanding of the function of time in speech is to take a closer look at acoustic segment phenomena (as was done in Paper 3). One such example is the c-center effect (Browman & Goldstein, 1988; Byrd, 1995). This phenomenon has not been tested in the individual studies but is still addressed in Papers 1 and 2; it has also been a major source of inspiration for the dissertation. However, the questions posed in the introduction about the c-center effect still remain unanswered. In short: does the c-center effect also occur in the gestural onset-onset relationship, or is it just a target phenomenon?

The proposed posture acceleration intervals of the consonants may have a one-to-one relation to the realization of the segment. Thus, the center of the consonant, the c-center, is found within the posture interval. A c-center effect should thus arise as a result of the different postures being shifted either way, which in other words is the same as the posture onset, maximal deceleration, having a different timing between consonants. However, the deceleration having a different timing does not necessarily mean that the onsets of the precedent active acceleration interval also have a different timing. They may still have simultaneous onsets while the active acceleration intervals have different lengths. Active acceleration intervals can have different lengths because their length is determined by speed, distance and the complexity of the target, as evidenced by consonant duration variations.

One possible way to take a closer look at the c-center effect is to investigate not only acceleration and deceleration, but also peak velocity. If peak velocity turns out to differ between consonants in a cluster, it is possible that the consonants move at different speed. In turn, this can lead to them reaching the posture at different points in time. Thus, one consonant is realized earlier than another simply because it is faster; in other words, a c-center effect occurs.

A deeper understanding of the c-center effect may bring clarity regarding the existence of an external clock or not. If gestural onsets are indeed simultaneous, the external clock might predetermine this, while the c-center effect that occurs at the phonemic target may be due to the inherent speed of the various articulators, which in turn is due to their tenseness. Such a connection would make it possible for speech at one level to be controlled by the external time structure, while at another level it would not.

5.2 The Swedish word accents

Linking the articulation of consonants and vowels with the tonal output may be too great and awkward a task for a dissertation. Undoubtedly, many more studies are needed to find an approach. In the discussion in Chapter 4, I make an attempt, where the f_0 change is suggested to be time-locked with articulatory movements, which are different depending on when in the word they occur and also whether the change is a f_0 rise or a f_0 fall. It is a hypothetical solution, which includes the auxiliary hypothesis that SWA consist of rising and falling tone gestures, and where you must first test and determine what in the f_0 change may be linked. In a falling tone gesture, is the whole larynx moved? And in such a case, how can we see it from the jaw movement and separate it from the mechanical movements that constitute a syllable? If it is a f_0 rise, of all the parameters that contribute to it, then what is time-locked, and with which articulator? Distinguishing between the mechanical and dynamic structures (that is, the measurable and the phenomena that underlie them) seems essential, but how can we do that?

If we look at the individual tones in SWA, we thus need to make new f_0 analyses of the data. The results of these analyses can in turn show that among our speakers there are not only hidden intra-speaker variations depending on the level of prominence (as proposed in Chapter 2, Methods), but very possibly also different dialect variants. As Meyer's data show (see e.g. Öhman, 1967), the Scania region shows various tonal patterns. The biggest difference seems to be between the northern and southern regions of Scania (Öhman, 1967), a distinction not made in this dissertation.

It is obvious that many questions remain. But, if we can solve some of them, we can start to look a little more at what distinguishes the Swedish dialects. If word accents really activate a f_0 fall, then what constitutes the f_0 rise? Does the rise actually consist of prominence after all? What determines when the fall takes place and what is it connected to? How can the timing in such cases differ between dialects, and why are there all of a sudden two peaks in one dialect but not in another? Maybe the dialect typology can even guide us here. If the tonal rise is allowed to vary systematically in time in the word between dialects, then that systematization is not lexical while, on the other hand, the difference between an early (read: A1) and a late (read: A2) tonal fall is lexical. Thus, a relationship that defines which and when specific laryngeal structures are timed with a specific articulatory movement seems to be an essential part of determining whether there is lexical difference or not.

Another possible path may be to continue analyzing the acceleration intervals to find out how the timing of f_0 changes is determined. If a time-locked relationship does not exist with either level in a presumed articulatory hierarchy, then how can the tones be realized so systematically? There can be no doubt that they are time-locked with articulation. In fact, Öhman's intonation model, the pulse model (Öhman, 1967), could once again be addressed. If the model is correct, which postulates that word accents are actually negative pulses (similar to the tonal fall) superimposed on tonal rises, we can, as far as South Swedish is concerned, determine that timing of A1 seems to take place at the end of C1 segment boundary. In other words, it occurs (with time delay included) at the same time as the acceleration at consonant offset, i.e. the onset of the active acceleration interval. Are we dealing with a negative pulse (according to Öhman) constituting the word accent, a pulse that occurs at the same time as a positive force is applied to an articulator to create acceleration? We may be witnessing a timing of different types of forces. Öhman's intonation model (1967) and the idea of different overlapping pulses for the raising and lowering of f_0, are of great interest in this context and should be re-evaluated.

5.3 Other implications and unresolved issues

5.3.1 Speech motor control

The studies in this thesis, as well as the discussion, have addressed prior research on speech motor control, or to motor control in general, to a very small extent. Without a doubt, there are aspects of motor control that I am not aware of, or that may have been represented in too simplified a manner in the dissertation. Either way, phonological models, not to mention models involving multimodality, have everything to gain from a knowledge of motor skills.

This thesis has not touched on how, for example, acceleration intervals work with internal models, motor commands, or sensorimotor feedback. However, a natural next step would be to determine what governs the applied force, which in theory affects how a damped mass-spring system of a specific articulator acts. Where does acceleration fit in? How are we able to control external force? The question of what targets are is closely linked to when actions are initiated, since speed is determined by what is to be performed. Some theories suggest virtual targets, i.e. targets that are not within the oral cavity (Löfqvist, 2005; Brunner et al., 2011). However, this is more suited to consonants and can describe the physical contact between different articulators (Löfqvist, 2005). However, how acceleration and deceleration are involved when reaching a target outside the oral cavity is unclear, since in a mass-spring system deceleration occurs as a result of the target being already reached. Perhaps several systems are involved, as well as several different targets during a consonant's articulation (see e.g. *via-points* in Perrier, 2012). On the other hand, because speech is so fast, and we have practiced it throughout our lives, the well-oiled speech apparatus might be more adapted to being started than to adjusting to different targets, whether in or outside the oral cavity. No matter what the explanation is, speech must consist of the most efficient and the least energy intensive systems.

5.3.2 Understanding speech disorders

In this dissertation, I have assumed that intra-personal variation is large, given the individual differences in the speech apparatus. Furthermore, I have assumed that inter-personal variation is also great but that it is of a systematic nature with context playing a key role. Thus, the systematic differences occur in all speakers but may act in various ways between speakers. It is a key question to figure out what can vary and what is systematic across speakers. Such a systematic building block seems to be that articulators must be timed with each other at given moments.

In general, timing appears to be an essential component of a functional phonotactic structure. The gestures need to occur in a specific order, and the movements of the articulators need to be timed with each other at certain landmarks (e.g. via-points) during the movement. If time exhibits a malfunction at a certain level in the speech production process, it will have an effect on the resulting speech signal. Depending on the level of the speech production process, this would supposedly yield different results. A malfunction in time on the selection of gestures results in a phonematic structure that does not match the intended lexical word, i.e. a speech error whereas a malfunction in time on, for example, the timing of the tongue tip may give the result that the target is not reached at the right time so that the sound "disappears" in the acoustic speech signal.

This brief overview of the structure of speech is intended as a narrative offspring and motivation for increased understanding of speech difficulties and disorders. Being understood by our fellow human beings is a fundamental value in life. It would be an honour to be able to contribute to a better understanding of the structures that need to be in place, and thus to be able to help those who lack one or two of these building blocks. As a researcher, one can easily get lost in all kinds of linguistic variation that can be measured and, hopefully, understood. It appears to be vital to keep an eye on the prize, which can be summed up as: what parts of speech need to work in coordination in order for speech to be intelligible?

5.3.3 What do the results mean for future perceptual research?

The dissertation began with an exposé of what the relationship might look like between articulation, acoustics and perception. This last concluding section intends to return to the point of view of perception, and asks itself: how can we move forward? The most obvious thought that strikes one first is that we should not start from the phonological toolbox we have today, with e.g. segments and phonemes. But, are there really other ways? Yes. If one assumes that speech is a dynamic structure of articulatory movements, movements that relate to each other in time, then we should be able to imagine that perception works on the basis of the same premise.

From a direct realist point of view, perceptual objects are phonological gestures of the vocal tract (Fowler, 1996). Or as Carol Fowler puts it:

> The acoustic signal is, after all, what the ear transduces; ears do not transduce articulations. The theories do not disagree on this point; they disagree on what the acoustic signal counts as for the perceiver. For acoustic theorists, it counts as a perceptual object; for me it counts as a specifier of speech events. (Fowler, 1996, p. 1737)

Hence, we undoubtedly hear the acoustics, but the big question is what speech events the acoustics convey.

The conundrum is that there is no consensus on the structure of the speech events. What we do know, however, is that as biological beings we use various signals to function in the outside world. Animals already use the relationship between differential equations (i.e. dynamical systems) as a prediction in perceptual systems in their physical environment (Iskarous, 2016). We work in the same way. When we hear a sharp sound, we know that there has been a "sharp" movement: distinct movements produce distinct sounds (Fowler, 1996). Thus, we can read what the movement is on the basis of the acoustics, because we are surrounded by dynamical systems and act on knowledge of these.

However, some kind of systematic movement pattern must exist and function in order for us to understand each other (i.e. a phonological system). This dissertation's results indicate that there is more systematic articulation in time than has been previously known and that we should examine further. In my work so far, I have only measured the mechanical movements, and not the phenomena behind them. Still, in mechanics, there seem to be tendencies of a basic dichotomous structure based on timing. For example, one based on the acceleration of movements. That is, one that is not only about the contrasts we usually think of in articulation, but a bimodal contrast in, for example, an early or late timing. What could be a better basis for a phonology than such a structure? That is, a basic dichotomous structure found in the timing of articulation.

Populärvetenskaplig sammanfattning på svenska

Man kan tro att vi pratar som vi skriver: att vi bygger ihop ord med hjälp av bokstäver som vi formar med munnen. Det är både sant och inte sant.

Vi bygger ihop ord, men inte med hjälp av bokstäver utan med hjälp av språkljud. Och vi bygger inte bara på längden, utan också på höjden. Tänk er olika stora lego-bitar: tvåor, fyror, åttor och tior. De långa bitarna (åttorna och tiorna) byggs efter varandra i rad på en platta. Ovanpå byggs en rad till med de kortare bitarna (tvåor och fyror), men med lite mellanrum emellan dem. Legobitarna överlappar varandra, men bara delvis. På samma sätt överlappar språkljuden varandra. De rörelser vi gör med munnen för att skapa vokalerna är långa och följer på varandra. Konsonanternas rörelser är snabba och korta och de ligger som ett lager ovanpå vokalerna. De kan både börja och sluta vid olika tidpunkter. Vissa av rörelserna startar dessutom samtidigt och går därför knappt att skilja åt. På senare år har man föreslagit att rörelserna består av *gester*, och att gester, inte språkljud, är talets minsta enheter.

I min avhandling undersöker jag hur gester påverkas av olika faktorer, som olika konsonantkombinationer eller melodimönster (som i svenska ordaccenter). Jag har spelat in ca 20 skåningar med hjälp av en *artikulograf*. Man fäster sensorer på ansiktet (t.ex. läpparna, näsan) och inuti munnen (t.ex. tandraden, tungspetsen, tungan). När personen sedan pratar i artikulografen fångar ett elektromagnetiskt fält upp hur sensorerna rör sig i tid och rum. Denna metod är bra för att ta reda på när gester (d.v.s. rörelser) startar och slutar, och vilka gester som ”vill vara tillsammans”. Rörelsestudierna har t.ex. visat att tunga, läppar och käke rör sig på olika vis beroende på om melodin i ordet börjar med en hög ton eller en låg.

I avhandlingen undersöker jag även hur gesterna styrs av tid: När startar en gest? Hur formas artikulationen av olika viktiga tidpunkter? Jag visar att man kan dela upp gester i tillstånd av aktiv rörelse och tillstånd av någorlunda stillhet eller positionering. Tänk er en talare (Anna) som skapar ett *m*: Annas läppar rör sig på en given signal mot ett mål. Läpparna bromsar in för att vara slutna en stund tills en signal säger att de ska återgå. Detta sker väldigt snabbt, och den slutna stunden i *m*:et är mycket kort, kanske 100 ms. Denna läpp-position är det man brukar mena med språkljudet *m.* Men egentligen är ett *m*, från start till slut, mycket längre än så. Det startar redan när läpparna får den första signalen, och avslutas efter att läpparna öppnats igen.

Avhandlingen visar att gesterna som bygger upp ett ord samverkar med varandra. Till exempel är läpparna i *m* slutna kortare tid om nästa konsonant också är ett *m*, som i ordet *mamma*. Med hjälp av hur *m*:et låter skulle en lyssnare därför kunna lista ut vilka de övriga konsonanterna i ordet är. Detta kan vara en av förklaringarna till att vi så snabbt kan förstå varandra i ett vanligt samtal.

Att studera gester är viktig grundforskning som är till stor hjälp för annan språkforskning. Jag vill särskilt lyfta fram vikten av tid för hur talet i slutändan låter, både i egenskap av gesternas timing och deras längd.

References

Adobe Inc. (2019). *Adobe Illustrator*. Retrieved from https://adobe.com/products/illustrator

Adobe Inc. (2019). *Adobe Indesign*. Retrieved from https://adobe.com/products/indesign

Ambrazaitis, G., & Bruce, G. (2006). Perception of South Swedish Word Accents. *Working papers, Lund University, Department of Linguistics and Phonetics, 52*, 5–8.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67(1), 1–48.

Beddor, P. S., McGowan, K. B., Boland, J. E., Coetzee, A. W. & Brasher, A. (2013). The time course of perception of coarticulation. *The Journal of the Acoustical Society of America,* 133, 2350–2366. doi:10.1121/1.4794366

Bell-Berti, F. & Harris, K. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. *The Journal of the Acoustical Society of America* 65, 1268–1270, doi:10.1121/1.382794

Boersma, P., Weenink, D. (2018). *Praat: doing phonetics by computer.* [Computer program] Version 6.0.37. Retrieved 3 February 2018 from http://www.praat.org/

Bombien, L., Mooshammer, C., & Hoole, P. (2013). Articulatory coordination in word-initial clusters of German. *Journal of Phonetics, 41*(6), 546–561. doi:10.1016/j.wocn.2013.07.006

Bootsma, R.J., Fernandez, L. & Mottet, D. (2004). Behind Fitts' law: kinematic patterns in goal-directed movements. *Int. K Human-Computer Studies* 61, 811–821.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219–252.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica, 45*(2–4), 140–155. doi:10.1159/000261823

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6*, 201–251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica, 49*(3–4), 155–180. doi:10.1159/000261913

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Bruce, G. (2007). Components of a prosodic typology of Swedish intonation. In T. Riad, C. Gussenhoven (Eds.), *Tones and Tunes – Volume 1: Typological Studies in Word and Sentence Prosody*, (pp. 113–146). Mouton de Gruyter, Berlin; New York.

Bruce, G. & Engstrand, O. (2006). The phonetic profile of Swedish. *Sprachtypologie und Universalienforschung*, 12–35.

Brunner, J., Fuchs, S. & Perrier, P. (2011). Supralaryngeal control in Korean velar stops. *Journal of Phonetics* 39, 178–195. doi:10.1016/j.wocn.2011.01.003

Byrd, D. (1995). C-Centers revisited. *Phonetica, 52*, 285–306. doi:10.1159/00026218

Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005). Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America, 118*(6), 3860–3873. doi:10.1121/1.2130950

Cho, T. (2002). *The effects of prosody on articulation in English*. New York: Routledge.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words.* Cambridge, MA: MIT Press.

D'Imperio, M., Espesser, R., Lœvenbruck, H., Menezes, C., Nguyen, N. & Welby, P. (2007). Are tones aligned with articulatory events? Evidence from Italian and French. In: Cole, J., Hualde, J. (Eds.), *Papers in Laboratory Phonology IX: Change in Phonology* (pp. 577–608). Mouton de Gruyter, The Hague.

Elert, C.-C (1964). *Phonologic Studies of Quantity in Swedish*. Almqvist & Wiksell, Uppsala.

Erickson D. (1977). The function of strap muscles in speech: pitch lowering or jaw opening. *Haskins Lab Status Rep Speech Res*, SR-49, 97–102.

Erickson, D., Baer, T. & Harris, K. (1983). The role of the strap muscles in pitch lowering. In D.M. Bless & J.H. Abbs (eds.), *Vocal fold physiology: Contemporary research and clinical issue* (pp. 279–285). College-Hill Press, San Diego, CA.

Erickson, D. (1998). Effects of contrastive emphasis on jaw opening. *Phonetica*, 55, 147–169.

Erickson, D., Iwata, R., Endo, M. & Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. *Proc. Tonal Aspects of Languages* Beijing, 53-56.

Erickson, D., Kawahara, S., Moore, J., Menezes, C., Suemitsu, A., Kim, J., & Shibuya, Y. (2014). Calculating articulatory syllable duration and phrase boundaries. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mu̇cke (Eds.), *Proceedings of the 10th International Seminar on Speech Production (ISSP), Cologne, Germany 2014*, 102-105.

Fant, G. & Lindblom, B. (1961). Studies of minimal speech sound units. *Speech Transmission Laboratory: Quarterly Progress Report, 2,* 1–11.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realistic perspective. *Journal of Phonetics, 14,* 3–28.

Fowler, C. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America* 99, 1730; doi:10.1121/1.415237

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36*(2–3), 171–195. doi:10.1177/002383099303600304

Frid, J., Svensson Lundmark, M., Ambrazaitis, G., Schötz, S., & House, D. (2019) Investigating visual prosody using articulography. In *Proceedings of 4th Conference of The Association Digital Humanities in the Nordic Countries: Copenhagen, March 6-8 2019* CEUR.

Fujimura, O. (2000). The C/D model and prosodic control of articulatory behaviour. *Phonetica*, 57, 128–138.

Gao, M. (2008). Tonal alignment in Mandarin Chinese: An articulatory phonology account. (Unpublished doctoral dissertation). Yale University, New Haven.

Garlén C (1988) *Svenskans fonologi*. Studentlitteratur, Lund.

Goldstein L (2019) The Role of Temporal Modulation in Sensorimotor Interaction. *Front. Psychol.* 10:2608. doi:10.3389/fpsyg.2019.02608

Goldstein, L. & Fowler, C. (2003). Articulatory phonology: A phonology for public language use. In N. O. Schiller & A. Meyer (Eds.), *Phonetics and phonology in language comprehension and production: Differences and similarities,* (pp. 159–207). Berlin: Mouton de Gruyter.

Gracco, V. (1988). Timing factors in the coordination of speech movements. *The Journal of Neuroscience, 8,* 4628–4639.

Gårding, E. (1967). *Internal juncture in Swedish*. Lund: Gleerup.

Gårding, E. & Lindblad, P. (1973). Constancy and variation in Swedish word accent patterns. *Working Papers, Phonetics laboratory, Lund University* 7, 36–110.

Gårding, E. (1974). Kontrastiv prosodi. Lund: Gleerup.

Gårding, E., Fujimura, O., Hirose, H., & Simada, Z. (1975). Laryngeal control of Swedish word accents. *Working papers, Lund University, Department of Linguistics and Phonetics, 10*, 53–82.

Hacking, I. [1983] (2010). *Representing and intervening. Introductory topics in the philosophy of science*. New York: Cambridge University Press.

Hall, N. (2010). Articulatory Phonology, *Language and Linguistic's Compass 4/9.*

Harrington, J. (2013). Acoustic Phonetics. In *The handbook of phonetic sciences*/edited by William J. Hardcastle, John Laver, Fiona E. Gibbon. – 2nd ed, (pp. 81–129). Wiley Blackwell.

Hirose, H. (2013). Investigating the physiology of laryngeal structures. In *The handbook of phonetic sciences*/edited by William J. Hardcastle, John Laver, Fiona E. Gibbon. – 2nd ed, (pp. 130–152). Wiley Blackwell.

Honda, K., Hirai, H., Masaki, S. & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in f0 control. *Lang. Speech* 42(4), 401–411.

Hoole, P., Bombien, L., Kühnert, B., and Mooshammer, C. (2009). Intrinsic and prosodic effects on articulatory coordination in initial consonant clusters. In *Frontiers in Phonetics and Speech Science*, edited by G. Fant, H. Fujisaki, and J. Shen, (pp. 275–286). The Commercial Press, Beijing.

Hoole, P. & Zierdt, A. (2010). Five-dimensional articulography. In: *Speech Motor Control: New developments in basic and applied research*, Editors: Ben Maassen, Pascal H.H.M. van Lieshout, (pp. 331–349). OUP.

Hoole, P. (2014). Recent work on EMA methods at IPS Munich. Retrieved from https://www.phonetik.uni-muenchen.de/~hoole/articmanual/ag501/carstens_workshop_summary_issp2014.pdf

Iskarous, K., Fowler, C. & Whalen, D. (2010). Locus equations are an acoustic expression of articulator synergy. *The Journal of the Acoustical Society of America* 128, 2021; doi:10.1121/1.3479538

Iskarous, K. (2016). Compatible Dynamical Models of Environmental, Sensory, and Perceptual Systems, *Ecological Psychology*, 28:4, 295–311. doi:10.1080/10407413.2016.1230377

Iskarous, K. (2017). The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics* 64, 8–20.

Jakobson, R. & Halle, M. (1956). *Fundamentals of language*. The Hague: Mouton.

Jakobson, R. Fant, G. & Halle, M. (1969). *Preliminaries to speech analysis*. 8th printing. Cambridge: MIT Press

Katsika, A, Krivopapić, J., Moonshammer, C., Tiede, M. & Goldstein, L. (2014). The coordination of boundary tones and its interaction with prominence. *Journal of Phonetics,* vol. 44, 62–82.

Kawahara, S., Masuda, H., Erickson, D., Moore, J., Suemitsu, A., & Shibuya, Y. (2014). Quantifying the effects of vowel quality and preceding consonants on jaw displacement: Japanese data. *Journal of the Phonetic Society of Japan, 18*(2), 54–62. doi:10.24467/onseikenkyu.18.2_54

Kawato, M., Maeda, Y., Uno, Y. & Suzuki, R. (1990). Trajectory formation of arm movement by cascade neural network model based on minimum torque-change criterion. *Biological Cybernetics*, 62, 275–288.

Kent, R. (1997). Gestural Phonology: Basic Concepts and Applications in Speech-Language Pathology. In M. J. Ball & R. D. Kent (Eds.), *The new phonologies: developments in clinical linguistics*, San Diego: Singular, 247–268.

Krivokapić, J., Tiede, M. K., & Tyrone, M. E. (2017). A kinematic study of prosodic structure in articulatory and manual gestures: Results from a novel method of data collection. *Laboratory Phonology, 8,* 3.

Kroos, C., Hoole, P., Kühnert, B. & Tillmann, H. (1997). Phonetic evidence for the phonological status of the tense-lax distinction in German. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 35*, 17–25.

Kühnert, B, Hoole, P. & Mooshammer, C. (2006). Gestural overlap and C-center in selected French consonant clusters. *Proc. 7th Int. Seminar on Speech Production,* Ubatuba, 327–334.

Kuznetsova, A., Brockhoff, P. & Christensen, R. (2017). lmerTest Package: tests in linear mixed effects models. *J. Stat. Software* 82(13), 1–26. doi:10.18637/jss.v082.i13

Lindblom, B. (1963). Spectrographic study of vowel reduction. *The Journal of the Acoustical Society of America* 35, 1773–1781.

Lindblom, B. (1983). Economy of speech gestures. In MacNeilage, P. (ed) *The Production of Speech*, (pp. 217–245). New York: Springer-Verlag.

Lindblom, B. & Maddieson, I. (1988). Phonetic universals in consonant systems. In: *Language, speech and mind,* ed. L. M. Hyman & C. N. Li. Routledge.

Löfqvist, A. (1975). Intrinsic and extrinsic f0 variations in Swedish tonal accents. *Phonetica, 31*, 228–247. doi:10.1159/000259671

Löfqvist, A. (2003). Interarticulator programming in speech production. *Proceedings of the 15th International Congress of Phonetic Sciences*. Barcelona, Spain, 27–32.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. *The Journal of the Acoustical Society of America* 117, 858. doi:10.1121/1.1840531

Löfqvist, A. (2006). Interarticulator programming: Effects of closure duration on lip and tongue coordination in Japanese. *The Journal of the Acoustical Society of America* 120, 2872. doi:10.1121/1.2345832

Löfqvist, A. (2007). Tongue movement kinematics in long and short Japanese consonants, *The Journal of the Acoustical Society of America* 122 (1), 512–518.

Löfqvist, A. & Gracco, V. (1999). Interarticulator programming in VCV sequences: Lip and tongue movements. *The Journal of the Acoustical Society of America* 105, 1864 –1876. doi:10.1121/1.426723

Löfqvist, A. & Gracco, V. (2002). Control of oral closure in lingual stop consonant production. *Journal of the Acoustical Society of America*, 111(6), 2811–2827.

Macneilage, P. F., & Declerk, J. L. (1969). On motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America, 45*(5), 1217–1213. doi:10.1121/1.1911593

Marin, S. (2013). The temporal organization of complex onsets and codas in Romanian: A gestural approach. *Journal of Phonetics, 41*(3–4), 211–227. doi:10.1016/j.wocn.2013.02.001

Marin, S. & Pouplier, M. (2104). Articulatory synergies in the temporal organization of liquid clusters in Romanian. *Journal of Phonetics 42*, 24–36.

Marslen-Wilson, W. & Zwitserlood, P. (1989) Accessing Spoken Words: The Importance of Word Onsets. In *Journal of Experimental Psychology: Human Perception and Performance*. Vol. 15, no. 3, 576–58.

Moen, I. (2006). Analysis of a case of the foreign accent syndrome in terms of the framework of gestural phonology. *Journal of Neurolinguistics*, 19, 410–423.

Mooshammer, C., Hoole, P., & Geumann, A. (2006). Interarticulator cohesion within coronal consonant production. *Journal of the Acoustical Society of America, 120*(2), 1028–1039. doi:10.1121/1.2208430

Mooshammer, C., Hoole, P. & Geumann, A. (2007). Jaw and order. *Lang. Speech* 50(2), 145–176.

Mooshammer, C., Bombien, L., & Krivokapic, J. (2013). Prosodic effects on speech gestures: A shape analysis based on functional data analysis. *The Journal of the Acoustical Society of America, 133*, 3565. doi:10.1121/1.4806505

Mücke, D., Nam, H., Hermes, A., & Goldstein, L. M. (2012). Coupling of tone and constriction gestures in pitch accents. In P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, & B. Kühnert (Eds.), *Consonant clusters and structural complexity*, (pp. 205–230). Berlin: Mouton de Gruyter.

Mücke, D. & Grice, M. (2014). The effect of focus marking on supralaryngeal articulation – Is it mediated by accentuation? *Journal of Phonetics* 44, 47–61.

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, E. Marsico, I. Chitoran, & C. Coupé (Eds.), *Approaches to phonological complexity*, (pp. 299-328). Berlin: Mouton de Gruyter.

Niemann, H., Grice, M. & Mücke, D. (2014). Segmental and positional effects in tonal alignment: An articulatory approach. *Proceedings of 10th ISSP, Cologne,* 285–288.

Niemann, H. & Mücke, D. (2015). Effects of phrasal position and metrical structure on alignment patterns of nuclear pitch accents in German: Acoustics and articulation. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK.

Ohala, J. J., & Eukel, B. W. (1987). Explaining the intrinsic pitch of vowels. In R. Channon, & L. Shockey (Eds.), *In honor of Ilse Lehiste*, (pp. 207–215). Dordrecht: Foris.

Pastätter, M. & Pouplier, M. (2017). Articulatory mechanisms underlying onset-vowel organization. *Journal of Phonetics*, 65, 1–14.

Perrier, P. (2012). Gesture planning integrating knowledge of the motor plant's dynamics: A literature review from motor control and speech motor control. In S. Fuchs, M. Weirich, D. Pape & P. Perrier (Eds.), *Speech Planning and Dynamics*, (pp.191–238). Peter Lang Publishers, Speech Production and Perception.

Popper, K. [1959] (2002). *The Logic of Scientific Discovery*. London and New York: Routledge Classics.

R Core Team. (2015). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

Recasens, D. (2002). An EMA study of VCV coarticulatory direction. *Journal of the Acoustical Society of America, 111*(6), 2828–2841. doi:10.1121/1.1479146

Riad, T. (2006). Scandinavian accent typology. *Sprachtypologie und Universalienforschung* 59, 36–55.

Riad, T. (2014). *The phonology of Swedish*. Oxford: Oxford University Press.

Roll, M., Söderström, P. &Horne, M. (2013). Word stem tones cue suffixes in the brain. *Brain Res.* 1520, 116–120. doi:10.1016/j.brainres.2013.05.013

Saltzman, E. L. & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1, (pp. 333–382).

Schaeffler, F. (2005). Phonological Quantity in Swedish Dialects: Typological aspects, phonetic variation and diachronic change. Ph.D. thesis, Umeå University, Umeå.

Schötz, S., Frid, J. & Löfqvist, A. (2013). Development of speech motor control: lip movement variability. *The Journal of the Acoustical Society of America*, 133, 4210–4217.

Shaw, J., Chen, W., Proctor, M. & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *J. Speech Lang. Hear. Res.* 59(6), 1566–1574.

Shaw, J., & Chen, W. (2019). Spatially-conditioned speech timing: evidence and implications. *Frontiers in Psychology, 10*, 2726. doi:10.3389/fpsyg.2019.02726

Sigurd, B. (1965). *Phonotactic structures in Swedish*. Lund: Uniskol.

Stone, M. (2013). Laboratory techniques for investigating speech articulation. In *The handbook of phonetic sciences*/edited by William J. Hardcastle, John Laver, Fiona E. Gibbon. – 2nd ed, (pp. 9–38). Wiley Blackwell.

Svensson Lundmark, M. (2018). Durational properties of word-initial consonants – an acoustic and articulatory study of intra-syllabic relations in a pitch-accent language. In *Proceedings of Fonetik 2018,* Gothenburg, Sweden, 65–66.

Svensson Lundmark, M., & Frid, J. (2018). Word onset CV coarticulation affected by post-vocalic consonants. Poster session presented at LabPhon16, Lisbon, Portugal.

Svensson Lundmark, M., Frid, J., & Schötz, S. (2015). A pilot study: acoustic and articulatory data on tonal alignment in Swedish word accents. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK, Paper number 590.

Tilsen, S. (2016). Selection and coordination: The articulatory basis for the emergence of phonological structure. *Journal of Phonetics,* 55, 53–77.

Turco, G. & Braun, B. (2016). An acoustic study on non- local anticipatory effects of Italian length contrast. *J. Acoust. Soc. Am.* 140, 2247–2256. doi:10.1121/1.4962982

Turk, A. & Shattuck-Hufnagel, S. (2020). Timing Evidence for Symbolic Phonological Representations and Phonology-Extrinsic Timing in Speech Production. *Front. Psychol.* 10:2952. doi:10.3389/fpsyg.2019.02952

Türk, H., Lippus, P. & Simko, J. (2017). Context-dependent articulation of consonant gemination in Estonian. *Laboratory Phonology: Journal of the Association for Laboratory Phonology 8(1),* 1–26.

Wedel, A., Ussishkin, A. & King, A. (2019). Incremental word processing influences the evolution of phonotactic patterns. *Folia Linguistica Historica 40(1),* 231–248. doi:10.1515/flih-2019-0011

Wieling, M., & Tiede, M. (2017). Quantitative identification of dialect-specific articulatory settings. *Journal of the Acoustical Society of America,* 142(1), 389–394. doi:10.1121/1.4990951

Yi, H. & Tilsen, S. (2014) Interaction between lexical tone and intonation: an EMA study. *Proceedings of Interspeech 2016*, San Francisco, USA, 2448–2452.

Xu, Y. (2013). ProsodyPro — a tool for large-scale systematic prosody analysis. in *Proc. Tools and Resources for the Analysis of Speech Prosody 2013* Aix-en-Provence, 7–10.

Zsiga, E. (1994). Acoustic evidence for gestural overlap in consonant sequences, *Journal of Phonetics,* vol. 22, 121–140.

Öhman, S. (1965). On the coordination of articulatory and phonatory activity in the production of Swedish tonal accents. *Quarterly Progress and Status Report, Dept. for Speech, Music and Hearing, KTH Stockholm, 6*, 14–19.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America,* 39(1), 151–168. doi:10.1121/1.1909864

Öhman, S. (1967). Word and sentence intonation: A quantitative model. *Quarterly Progress and Status Report, Dept. for Speech, Music and Hearing, KTH Stockholm*, 8(2-3), 20–54.

Appendices

Appendix I: Speech material (pilot project)

1. Det var BILEN jag sa.
2. Det var BILAR jag sa.
3. Det var BILDEN jag sa.
4. Det var BILDER jag sa.
5. Det var BOVEN jag sa.
6. Det var BOVAR jag sa.
7. Det var VOVVEN jag sa.
8. Det var VOVVAR jag sa.
9. Det var VALEN jag sa.
10. Det var VALAR jag sa.
11. Det var VALLEN jag sa.
12. Det var VALLAR jag sa.

Appendix II: Speech material (corpus)

(In this version target words are underlined; the material presented on the screen to the speakers did not have underlined target words.)

1. Vilken häst kammade Jonas manen på?|Jonas kammade manen på Ronja.

2. Vad rimmade Isak manar med?|Isak rimmade manar med ”svanar”.

3. Hur rammade Eva mannen?|Eva rammade mannen med cykeln.

4. Hur skummade Inger manna?|Inger skummade manna med vispen.

5. Med vem kollade Sanna filmen ”Mammut”?|Sanna kollade ”Mammut” med Nora.

6. Var lämnade morfar mamma?|Morfar lämnade mamma med doktorn.

7. Vad rimmade Eskil ”nanny” med?|Eskil rimmade ”nanny” med ”Fanny”.

8. Hur kammade Tareq Nanna?|Tareq kammade Nanna med kammen.

9. Vem lämnade pappa namnen till?|Pappa lämnade namnen till vakten.

10. Vad rimmade Lina ”namnar” med?|Lina rimmade ”namnar” med ”famnar”.

11. Var mumlade Elin Amen?|Elin mumlade Amen hos prästen.

12. Vad rimmade Leila amma med?|Leila rimmade amma med ”damma”.

13. Vad rammade Lisa bilen med?|Lisa rammade bilen med cykeln.

14. Vem limmade Anna bilar med?|Anna limmade bilar med morfar.

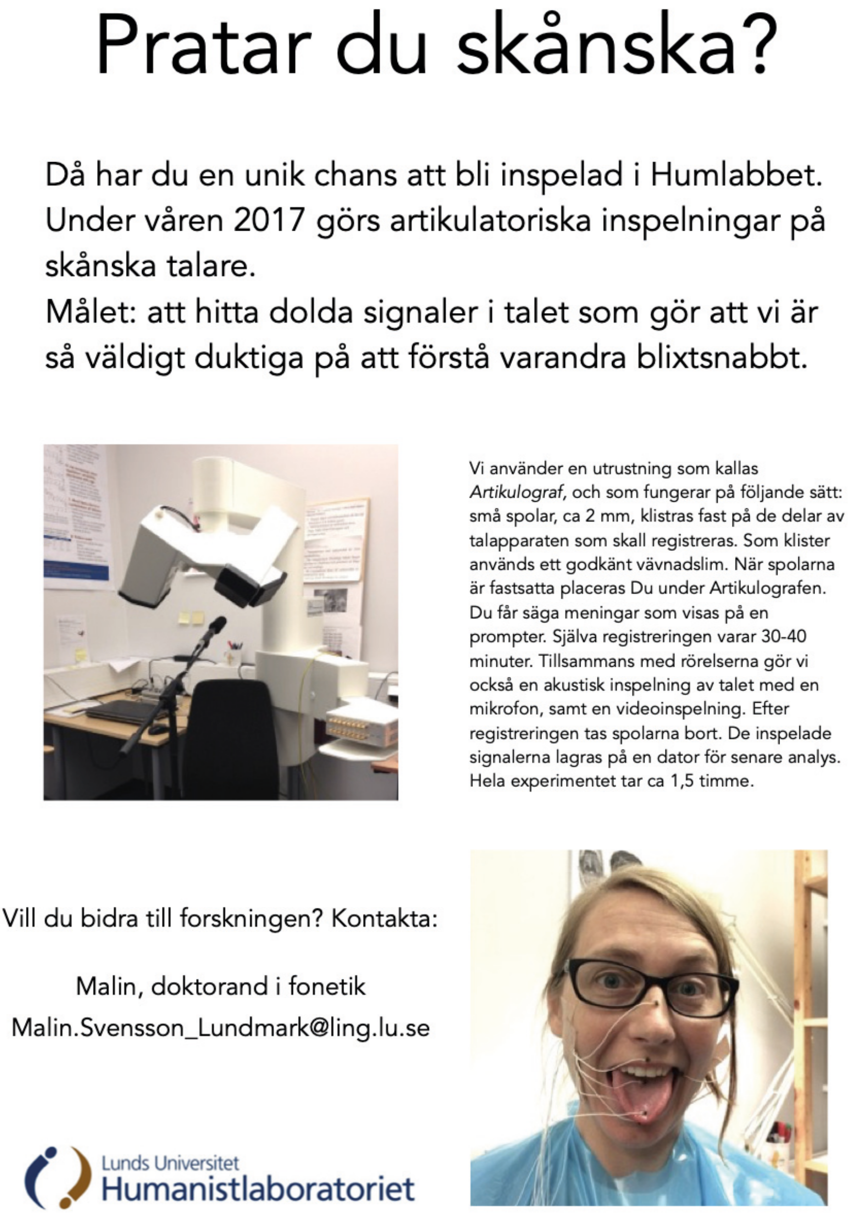
15. När lämnade Kalle balen?|Kalle lämnade balen vid midnatt.

16. Var lämnade bonden alla balar?|Bonden lämnade balar vid ladan.

17. Var lämnade Inger malen hon fångat?|Inger lämnade malen till kocken.

18. Var lämnade Åsa alla malar hon fångat?|Åsa lämnade malar vid båten.

Appendix III: Poster data collection, in Swedish (corpus)



Pratar du skånska?

Då har du en unik chans att bli inspelad i Humlabbet. Under våren 2017 görs artikulatoriska inspelningar på skånska talare.
Målet: att hitta dolda signaler i talet som gör att vi är så väldigt duktiga på att förstå varandra blixtsnabbt.

Vi använder en utrustning som kallas *Artikulograf*, och som fungerar på följande sätt: små spolar, ca 2 mm, klistras fast på de delar av talapparaten som skall registreras. Som klister används ett godkänt vävnadslim. När spolarna är fastsatta placeras Du under Artikulografen. Du får säga meningar som visas på en prompter. Själva registreringen varar 30-40 minuter. Tillsammans med rörelserna gör vi också en akustisk inspelning av talet med en mikrofon, samt en videoinspelning. Efter registreringen tas spolarna bort. De inspelade signalerna lagras på en dator för senare analys. Hela experimentet tar ca 1,5 timme.

Vill du bidra till forskningen? Kontakta:

Malin, doktorand i fonetik
Malin.Svensson_Lundmark@ling.lu.se

Lunds Universitet
Humanistlaboratoriet

Appendix IV: Speakers metadata (corpus)

SPEAKERS	BIRTHPLACE	RAISED	PARENTS
M1	Malmö	Malmö	Lund and Osby
M2	Lund	Bjärred, Häglinge (Hässleholm)	m: Kristianstad f: Malmö
M3	Lund	Lund	m: Halmstad f: Lund
F1	Ystad	Skurup	Österlen: f: Valleberga, m: Kåseberga
F2	Ystad	Ystad	Malmö
F3	Lund	Lund	m: Malmö f: London
M4	Helsingborg	Helsingborg	m: Mörarp (Helsingborg) f: Höganäs
F4	Ystad	Skurup	Österlen (m: Kåseberga f: Peppinge)
M5	Ystad	Malmö	m: Uppsala, f: Västervik
M6	Eslöv	Eslöv	m: Malmö, f: Österlen (västra)
F5	Helsingborg	Göinge (Farstorp)	Klippan
F6	Lund	Bjärred, Löddeköpinge	m: Lund, f: Linköping
M7	Tunnby (Tomelilla)	Tunnby (Tomelilla)	m: Sälshög (Tomelilla), f: Tunnby (Tomelilla)
M8	Malmö	Malmö	Malmö, Grandfather: Alnarp
F7	Lund	S Sandby, Revinge	m: utanför Lund f: Irland
M9	Billeberga (Svalöv)	Billeberga (Svalöv)	Öved, Bjärsjölagård
F8	Malmö	Malmö	m: Skanör, f: Danmark
F9	Lund	Lund	m: Fleninge f: Helsingborg
F10	Malmö	Vittsjö	m: Göteborg, Malmö. f: Malmö
F11	Lund	Landskrona	m: Landskrona f: Tyskland
F12	Vittsjö	Vittsjö	m: Göteborg, Malmö. f: Malmö

Abbreviations: F = female speaker, M = male speaker, f = father, m = mother

Attached papers

Paper I

INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents

Malin Svensson Lundmark, Gilbert Ambrazaitis, Otto Ewald

Centre for Languages and Literature, Lund University, Sweden
malin.svensson_lundmark@ling.lu.se, gilbert.ambrazaitis@ling.lu.se

Abstract

This study investigates acoustic and articulatory correlates of South Swedish word accents (Accent 1 vs. 2) – a tonal distinction traditionally associated with F0 timing. The study is motivated by previous findings on (i) the acoustic complexity of tonal prosody and (ii) tonal-articulatory interplay in other languages.

Acoustic and articulatory (EMA) data from two controlled experiments are reported (14 speakers in total; pilot EMA recordings with 2 speakers). Apart from the well-established F0 timing pattern, results of Experiment 1 reveal a longer duration of a post-stress consonant in Accent 2 than in Accent 1, a higher degree of creaky voice in Accent 1, as well as a deviant (two-peak) pitch pattern in Accent 2 for one of eight discourse conditions used in the experiment. Experiment 2 reveals an effect of word accent on vowel articulation, as the tongue body gesture target is reached earlier in Accent 2. It also suggests slight but (marginally) significant word-accent effects on word-initial gestural coordination, taking slightly different forms in the two speakers, as well as corresponding differences in word-initial formant patterns. Results are discussed concerning their potential perceptual relevance, as well as with reference to the c-center effect discussed within Articulatory Phonology.

Index Terms: speech production, pitch, lexical tone, voice quality, articulatory gestures, articulatory phonology, EMA

1. Introduction

This paper studies acoustic and articulatory manifestations of South Swedish word accents: a binary, phonological tonal distinction (Accent 1, Accent 2, henceforth A1 and A2). It thereby adds to a growing body of evidence arguing for a multidimensional nature of tonal prosody (cf. 1.1). To this end, we explore further possible phonetic cues beyond the timing of a (rise-)fall in fundamental frequency (F0), which is traditionally regarded as the distinction's primary phonetic correlate (cf. 1.1). In particular, we investigate patterns of F0, voice quality, durations, articulatory gestures, and formants.

1.1. Acoustic complexity of Swedish word accents

The Swedish word accent distinction has been recognized as a tonal, i.e. F0-related phenomenon since at least Meyer's early comprehensive dialect study [1,2], and F0 has been established as the primary cue for perception [3,4,5]. The F0-dinstinction has been analyzed in terms of timing of a (rising-) falling F0 gesture [6], a proposal which has been controversial when it comes to the Stockholm variety [7,8], but rather undisputed for the South Swedish variety studied in this paper [9,10] (cf. Fig. 2 for a display of typical F0-patterns).

Further (potential) phonetic correlates of the word accents beyond F0 have hardly attracted any attention, although duration and intensity have been revealed as secondary phonetic correlates already in [3]. In particular, for Stockholm Swedish, a longer duration has been observed for the stressed vowel in A1 than in A2, while the reversed pattern was attested for a post-stress consonant, i.e. a longer duration for A2 [11]. Assuming that tonal complexity would trigger a longer duration, we would expect a longer vowel in A2 in Elert's [11] data, because the Stockholm A2 surfaces as a two-peak F0 pattern. So what can, alternatively, explain Elert's findings? We suggest that it is not the tonal complexity *per se*, but the function of the tones involved that matters: it is the *focal accent* tone that causes lengthening. Following [6], this tone is realized in the stressed syllable in A1, but in the post-stress in A2, explaining the differential lengthening of stressed vowel (A1) and following consonant (A2) in Elert's [11] data.

A multidimensional view of tonal word accent encoding is well in line with an increasing body of evidence attesting acoustic complexity of tonal prosodic events in several other languages, often in connection with sentence-level intonation [12, 13, 14, 15].

1.2. The interplay of tonal and articulatory gestures

The framework of Articulatory Phonology [16, 17] has in recent years started to include prosodic information [18, 19, 20, 21, 22, 23]. More specifically tones have been proposed to represent articulatory gestures, i.e. tone gestures, comparable to consonantal and vocalic gestures [19, 21], and tonal alignment has been shown to be more stable relative to articulatory landmarks [21, 24, 25, 26] than to acoustic ones [27, 28, 29, 30, 31]. Lexical tone gestures in Mandarin have even been proposed to compete with consonantal gestures in onset [19]. This type of competitive coordination between articulatory gestures is the source for a phenomenon known as the *c-center effect,* where the start of consonantal gestures shifts depending on an inter-competitive relationship with the vowel [32]. Thus, if vocalic, consonantal and tone gestures were coordinated in onset we would find cues of this coordination pattern in the consonantal and vocalic gestures.

1.3. Hypotheses

Based on the background presented in the previous sections we hypothesize, in general, that South Swedish word accents exhibit further acoustic or articulatory correlated beyond F0 timing. For instance, given the different tonal timing patterns of A1 and A2 we assume that the coupling of tone gestures with consonantal and vocalic gestures would differ between the word accents. We expect to find cues of this coupling in the coordination pattern of the consonantal and vocalic gestures, e.g. a difference in c-center effect in onset.

http://dx.doi.org/10.21437/Interspeech.2017-1502

Another prediction based on our interpretation of [11] (cf. 1.1) is that, for South Swedish, *no* durational effect should occur for the vowel, as the focal accent, in this dialect, is produced through the word accent gesture, and within the stressed vowel for both A1 and A2; we might, however, still predict a longer post-stress consonant in A2, as the A2 pitch rise-fall crosses this consonant.

2. Method

Two experiments are presented, designed independently of each other, both investigating South Swedish word accent production. Apart from the same target word pair being used in both data collections – *bilen* ('the car') for A1, and *bilar* ('cars') for A2 – they differ – and complement each other – in several respects. Experiment 1 was originally designed to investigate word accent production as a function of discourse context (cf. 2.1). It involves 12 speakers (6 female, 6 male) and provides a relatively large amount of data (576 recorded tokens), however, limited to acoustic recordings. Experiment 2 is a pilot study involving 2 female speakers (39 recorded tokens), combining acoustic with articulatory recordings using Electromagnetic Articulography (EMA). The recordings from both experiments were acoustically segmented into consonants and vowels. Additionally, the occlusion phase of /b/ was segmented.

For experiment 1, the target words (A1, A2) were embedded in short carrier phrases such as 'yes, by car', and eight different conditions (discourse contexts) were created in order to elicit different readings of the test phrase, e.g. as an assertion, a confirmation, a correction, an exclamation etc. The context was presented in written form, and for some conditions, there was an additional auditory context question. Participant's read the contexts (quiet) and the target phrase (aloud) from a computer screen in an experimental studio.

In order to avoid unnecessary F0 analysis errors, F0 calculation was performed in the time-domain based on 'pulses' automatically determined by Praat [33] which we manually corrected using ProsodyPro [34].

In experiment 2 kinematic data was recorded in an Electromagnetic Articulograph (EMA, Carstens AG501, sampling rate 250 Hz) at the Lund University Humanities Lab. Sound was recorded simultaneously using an external condenser microphone (t.bone EM 9600). The target words were produced in the carrier phrase *Det var TARGET jag sa* ('It was TARGET I said'). Two female speakers of South Swedish (age 38 and 49) read the material ten times each. The sentences were shown on a prompter in a random order.

The consonantal gesture of the stressed syllable /bi:/ is a bilabial closure, and the vocalic gesture of /bi:/ is a palatal narrow. Hence, EMA data from sensors on the upper and the lower lips (=lip aperture), and on the tongue body, were collected and further processed in R [35] and normalized for head movements.

3. Results and discussion

3.1. Fundamental frequency (F0) (Experiment 1)

As for F0, we restrict this paper to a visual analysis of the general shape or patterning of F0, since correlates beyond F0 shall be in focus in the first place. Figure 1 displays mean F0 curves for A2, averaged across our 6 female speakers, separately for all eight discourse contexts involved, as an example. Intonational expressions due to sentence- or discourse-level functions are outside the scope of this paper, but we include this display for two reasons: First, it demonstrates the relative stability of tonal timing, despite discourse-induced variation in parameters such as peak shape, height, or range. It thus replicates and confirms what is known about the South Swedish word accents' F0 patterning (cf. 1.1).

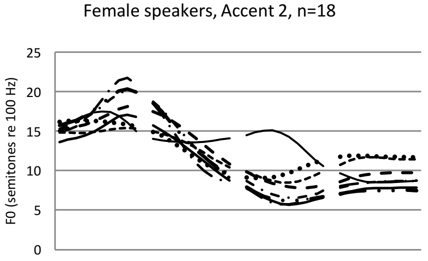


Figure 1: *Mean F0 curves across 6 speakers and 3 repetitions (n=18) for the target word (b)ilar 'cars', initial /b/ not included; 8 conditions (discourse contexts, represented by the separate lines); time is normalized, breaks in the curves indicate acoustic segment boundaries: /i/, /l/, /a/, /r/.*

Second, however, the data also provide an unexpected result, as the F0 pattern for one of the conditions (an 'exclamation') is crucially deviating; it indeed reminds of a Stockholm Swedish pattern, where A2 surfaces as a two-peak F0 curve. What is not evident from this mean curve is that the two-peak pattern was produced exclusively in this condition, in 13 of the 18 tokens, and at least once by each of the 6 speakers. We conclude that it represents a regular pattern of this dialect, occurring on certain discourse conditions, but will leave further discussion of this pattern for future research.

Figure 2 offers an average display comparing A1 and A2, across all female speakers and seven of the conditions (condition 'exclamation' excluded). We obtain equivalent results for male speakers.

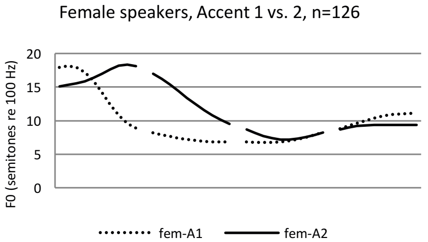


Figure 2: *Mean F0 curves across 6 speakers, 7 conditions (discourse contexts), and 3 repetitions (n=126) for the target words (b)ilen 'the car' (A1) and (b)ilar 'cars' (A2); for further explanations cf. Fig. 1.*

3.2. Voice quality (Experiment 1)

Informal observations during the annotation process revealed a high degree of creaky voice, mostly during the second (final) syllable of the target word (/lɛn/ or /lar/, respectively), but also, sometimes during the preceding (stressed) vowel /iː/. To study possible effects of the word accent on the occurrence of creaky voice, we annotated creaky voice (a) during the vowel /iː/ and (b) during the following consonant /l/, following a simple scheme deciding between absence/presence of creaky voice. Results are displayed in Table 1.

Table 1: *Annotations of creaky voice in %, broken down by word accent (A1, A2) and speaker gender.*

Condition	/iː/	/l/
A1 (all speakers)	62	79
A2 (all speakers)	1	54
A1 (female/ male)	74/50	78/79
A2 (female/ male)	1/1	63/44
Female/Male (both A)	38/25	71/62

Table 1 suggests a strong effect or word accent on the occurrence of creaky voice, despite a certain effect of gender, which is seen both in the stressed vowel and in the following voiced consonant. A linear mixed model for /iː/ with word accent and gender as fixed factors (speaker and context has random effects) reveals significant effects for word accent (*t*=-18.629, *df*=555, *p*<.001***), gender (*t*=-3.874, *df*=15.7, *p*=.0014**), and their interaction (*t*=4.180, *df*=555, *p*<.001***). For /l/, there is an effect of word accent (*t*=-3.290, *df*=555, *p*=.0011**), of the interaction of word accent and gender (*t*=-2.961, *df*=555, *p*=.0032**), but no main effect of gender.

A comparison of these findings with the F0 patterns obtained in 3.1 suggests that the occurrence of creaky voice relates to low or sharply falling F0, thus explaining its higher frequency of occurrence in A1, and in particular its occurrence already during the vowel, where it hardly ever occurs for A2.

3.3. Durations (Experiment 1 and 2)

Segmental durations of the stressed syllable and the following consonant (/b/, /iː/, /l/) were measured in the acoustic domain for data from both experiments. The results from Experiment 1 show no effect of word accent on the stressed syllable (neither on /b/ nor on /iː/), but a significant effect in the duration of the post-stress consonant /l/, which was on average 10 *ms* longer in A2 (89 *ms*) than in A1 (79 *ms*) (linear mixed model with word accent as fixed factor; speaker and context has random effects: *t*=8.90, *df*=556, *p*<.001***). This result is perfectly in line with our prediction formulated in 1.3. The results of Experiment 2, however, are less conclusive. They are inconsistent between the two speakers, and partly contradictory to our predictions. We explain these inconsistencies by the relatively small amount of data.

3.4. Articulatory gestures (Experiment 2)

In this section, we explore two different articulatory dimensions: (i) time lags of consonantal (bilabial closure, i.e. lip aperture, LA) and vocalic (palatal narrow, i.e. tongue body, TB) gesture onset at stressed-syllable onset (cf. 3.4.1), and (ii) timing of the target of the vocalic gesture (cf. 3.4.2). While (i) was motivated by the prediction of a c-center effect (cf. 1.3), (ii) was motivated by an initial qualitative assessment of the data, which suggest a later timing of TB gesture target in A1 than in A2.

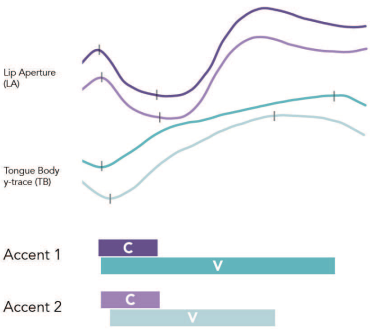


Figure 3: *Trajectories with marked onset and target of the consonantal gesture (bilabial closure, LA), and the vocalic gesture (palatal narrow, vertical position of TB), of an A1 and an A2 word.*

3.4.1. Gestural co-ordination in onset

Figure 3 above displays an example of articulatory gestural trajectories for an A1 token and an A2 token, as pronounced by speaker M. In this example, we see that the consonantal gesture starts slightly earlier in A2 than in A1. To corroborate this observation, we measured the time lags between onsets of the consonantal (LA) and the vocalic (TB y-trace) gestures. Results are displayed in Figure 4, suggesting for speaker M that the consonantal (LA) gesture starts somewhat earlier for A2 than for A1. This difference is significant for speaker M (*t*=-2.75, *df*=14.77, *p*=.015*), however, not for speaker S (*t*=-.44, *df*=18.91, *p*=.66; cf. Fig. 4).

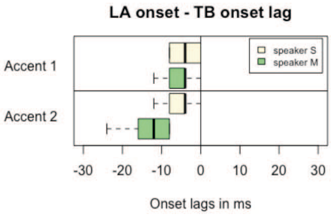


Figure 4: *Time lags of the consonantal gesture onset (LA) and the vocalic gesture onset (TB) in word onsets. Speaker M displays a small but significant c-center effect in Accent 2.*

An earlier onset of a consonantal gesture can be explained as the result of a c-center effect, when articulatory (consonantal or tonal, cf. 1.2 above) gestures co-occur and compete at syllable onset. In our case, it could be a result of (i) coordination with different tone gestures: presumably a H-gesture and a L-gesture, respectively. Studies on Mandarin have indeed reported different coordination patterns between the rising Tone 2 and the falling Tone 4 [19, 23]. Different coordination patterns could also be due to (ii) different coupling relations for A1 and A2. However, this is not within the scope of this article. An alternative explanation could assume (iii) a lexical tonal target in A2, but not in A1 [8], and relate the competitiveness to the phonological status of the tone. Indeed, post-lexical tone gestures have been shown to not result in a c-center effect, hence they seem to not compete with the consonantal gestures in onset [21].

Although the results do not suggest a c-center effect for speaker S, we observe traces of another, albeit related effect at syllable onset: For speaker S, the duration of the bilabial closure gesture (defined as the distance of LA target from LA onset) was on average slightly shorter in A2 than in A1 (marginally significant at *t*=1.97, *df*=11.79, *p*=.072), while no such effect did appear for speaker M (*t*=-.57, *df*=13.40, *p*=.58).

We suggest that these two effects observed for S and M might be two different outcomes of the same underlying mechanism, as they might have the equivalent effect of an earlier bilabial closure release (which need not, but might possibly following from a shorter LA gesture) with respect to the vocalic gesture. This would provide an earlier *acoustic* vowel onset in A2 than A1, which in turn might be perceptually motivated, supporting listeners anticipating the nature of the upcoming lexical accentual tone (a low tone in A2). An early recognition of lexical tones is advantageous as it has been shown for Swedish how word stem tones are used by listeners to predict upcoming word endings [36]. We will leave the verification of this perception-based interpretation to future research.

3.4.2. *Timing of the vocalic gestural target*

A salient articulatory difference between A1 and A2 consistently observed for both speakers was a later-timed gesture target (defined as the duration target–onset) of the vocalic gesture (i.e. the maximum height of the tongue body in /i/) in A1. This effect is seen in Figure 5; it is significant for both S (*t*=-5.84, *df*=14.28, *p*<.001***) and M (*t*=-2.53, *df*=11.33, *p*=.027*).

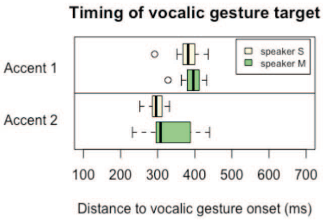


Figure 5: *Timing of the vocalic gesture target. For both speakers the vocalic gestural target (maximum height of the tongue body) is later in A1 than in A2.*

For a tentative interpretation of this effect we consider the possibility of a coupling to the durational effect attested in Experiment 1, as these two effects – a longer acoustic duration of the post-vocalic consonant (Exp. 1) and an earlier timing of the vocalic gesture target (Exp. 2) in A2 – are among the most consistent results of the respective experiments (after the expected results for F0). A longer acoustic duration of the consonant in A2 (Exp. 1) is possibly achieved by an earlier consonantal gesture; this in turn might trigger an earlier timing of the (pre-consonantal) vocalic gesture target. A partial support for this interpretation comes from additional measurements of the timing of the tongue tip gesture target for /l/, which indeed proved significantly earlier in A2 than in A1, however only for speaker S (*t*=-3.94, *df*=16.61, *p*=.0011**), but not for M (*t*=-.18, *df*=16.00, *p*=.86). However, an alternative interpretation is that the timing of the vocalic gesture target is very much related to the coupling of the tone gesture in A2, which could force the vocalic target timing to be earlier than in A1. A stable coordination pattern between the tone peak and the target of the vocalic gesture has been found in German pitch accents [26].

3.5. Formant patterns (Experiment 2)

If the articulatory findings reported in 3.4 are to be considered potential perceptual cues of Swedish word accents, then they should exhibit some acoustic correlate. We therefore measured the following formant frequencies: Following [37], who found F2 and F3 during stop closure to be affected by gestural overlap, we measured mean F2 and F3 of the occlusion phase of /b/, as well as F2 and F3 at vowel onset. The formants of the vowel onset were measured as a specific point immediately following the release of /b/. Two criteria were used to establish the point: 1) it was the second pulse with three distinct formants that constitute the vowel; 2) there was a steep rise in the amplitude at the vowel onset.

The results for the occlusion phase revealed an effect of word accent on the F3-F2 difference for speaker S (*t*=2.14, *df*=17.25, *p*=.047*), but not for F2 or F3 separately, and not at all for speaker M. For vowel onset, we found a complementary pattern: no effects of word accent for speaker S, but an effect on F3-F2 for M (*t*=2.83, *df*=12.62, *p*=.015*). That is, the c-center effect observed for M (cf. 3.4.1) relates to an acoustic effect at vowel onset, while the (marginal) difference in LA-duration in speaker S seems to relate to formant differences during the stop occlusion.

These results suggest that articulatory correlates of the word accents have acoustic effects and thus qualify as candidates for perceptual word accents cues.

4. Conclusion

Probably due to the strong attested power of F0 as a perceptual cue to the (South) Swedish word accent contrast (cf. 1), additional secondary cues, have so far hardly attracted any attention. This study suggests that South Swedish word accents are distinguished in speech production by means of several phonetic (articulatory and acoustic) dimensions: F0 timing, creaky voice predominantly in A1 , a longer duration of a post-stress consonant in A2, different gestural coordination at word onset (also mirroring in the acoustic domain in terms of the F3-F2 difference), and an earlier reached target of the vocalic gesture in A2. Future research will need to evaluate the perceptual relevance of these potential correlates of Swedish word accents and their usefulness, e.g., in the on-line prediction of upcoming suffixes [36].

5. Acknowledgements

We gratefully acknowledge the Humanities Laboratory, Lund University for offering experimental studios and EMA recording facilities. We also wish to thank Dr. Johan Frid (SWE-CLARIN) for assistance with data analysis (Exp. 2), as well as Anneliese Kelterer (student assistant) for annotating data from Exp. 1. Exp. 1 was conducted by the second author within a grant from the Swedish Research Council (PI Johan Frid, 2009-1566). Last but not least, we thank our participants!

6. References

[1] E. A. Meyer, “Die Intonation im Schwedischen,“ *Erster Teil: Die Sveamundarten*, Stockholm: Fritzes, 1937.

[2] B. Malmberg, “Bemerkungen zum schwedischen Wortakzent“, *Zeitschrift für Phonetik*, vol. 12, pp. 193–207, 1959.

[3] B. Malmberg, “Nyare fonetiska rön,” Lund: Gleerup, 1966.

[4] D. House and G. Bruce, “Word and focal accents in Swedish from a recognition perspective,” in *Nordic Prosody V – Papers from a Symposium, Turku,* pp. 156–173, 1990.

[5] G. Ambrazaitis and G. Bruce, “Perception of South Swedish Word Accents,” *Working Papers 52, Proceedings from Fonetik 2006*, G. Ambrazaitis, and S. Schötz, Eds. Lund, pp. 5–8, 2006.

[6] G. Bruce, “Swedish Word Accents in Sentence Perspective,” *Travaux de l'institut de linguistique de Lund*, no. 12, Lund: Gleerup, 1977.

[7] O. Engstrand, “Phonetic interpretation of the word accent contrast in Swedish,” *Phonetica*, vol. 52, pp. 171–179, 1995.

[8] T. Riad, “Scandinavian accent typology,” *Sprachtypologie und Universalienforschung (STUF),* vol. 59, no. 1, pp. 36–55, 2006.

[9] G. Bruce and E. Gårding, “A prosodic typology for Swedish dialects,” In *Nordic Prosody – Papers from a Symposium, Lund*, pp. 219–228, 1978.

[10] B. Malmberg, ”Sydsvensk ordaccent – en experimentalfonetisk undersökning,“ *Lunds Universitets Årsskrift. N. F. Avd. 1*, vol. 49, Lund: Gleerup, 1953.

[11] C.-C. Elert, “Phonologic Studies of Quantity in Swedish. Based on Material from Stockholm Speakers,” Stockholm: Almqvist & Wiksell, 1964.

[12] M. Heldner, “Focal accent – F0 movements and beyond”, *Ph. D. thesis*, Umeå University, 2001.

[13] V. Van Heuven and E. van Zanten, ”Speech rate as a secondary prosodic characteristic of polarity questions in three languages,” *Speech Communication* 47, pp. 87–99, 2005.

[14] O. Niebuhr, “The acoustic complexity of intonation,” *Nordic Prosody XI,* E-L. Asu and P. Lippus, Eds. Peter Lang, pp. 25–38, 2013.

[15] G. Ambrazaitis and J. Frid, “F0 peak timing, height, and shape as independent features,” *in Proc. 4th International Symposium on Tonal Aspects of Languages, Nijmegen,* pp. 138–142, 2014.

[16] C. Browman and L. Goldstein, “Towards an articulatory phonology,” *Phonology Yearbook 3*, pp. 219–252, 1986.

[17] C. Browman and L. Goldstein, “Articulatory phonology: An overview,” *Phonetica*, vol. 49, no. 3–4, pp. 155–180, 1992.

[18] D. Byrd and E. Saltzman, “The elastic phrase: modeling the dynamics of boundary-adjacent lengthening,” *Journal of Phonetics*, vol. 31, pp. 149–180, 2003.

[19] M. Gao, “Tonal alignment in Mandarin Chinese: An articulatory phonology account,” Doctoral dissertation, Yale University, 2008.

[20] H. Niemann, D. Mücke, H. Nam, L. Goldstein, and M. Grice, “Tones as Gestures: the Case of Italian and German,” in *Proceedings of ICPhS XVII, Hong Kong,* pp. 1486–1489, 2011.

[21] D. Mücke, H. Nam, A. Hermes, and L. Goldstein, “Coupling of tone and constriction gestures in pitch accents,” In *Consonant Clusters and Structural Complexity*, P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, and B. Kühnert, Eds. Munich: Mouton de Gruyter, pp. 157–176, 2012.

[22] A. Katsika, J. Krivopapic ́, C. Moonshammer, M. Tiede, and L. Goldstein, “The coordination of boundary tones and its interaction with prominence,” *Journal of Phonetics*, vol. 44, pp. 62–82, 2014.

[23] H. Yi and S. Tilsen, "Interaction between lexical tone and intonation: an EMA study," in INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, Sep 8–10, San Francisco, USA, Proceedings, pp. 2448–2452, 2016.

[24] Hermes, A., Becker, J., Mücke, D., Baumann, S., Grice, M. 2008. Articulatory Gestures and Focus Marking in German. *Proc. Speech Prosody 2008* Campinas, 457–460.

[25] H. Niemann, M. Grice, and D. Mücke, “Segmental and positional effects in tonal alignment: An articulatory approach,” *in Proceedings of 10th ISSP, Cologne,* pp. 285–288, 2014.

[26] H. Niemann and D. Mücke, “Effects of phrasal position and metrical structure on alignment patterns of nuclear pitch accents in German: Acoustics and articulation,” in *Proceedings of 18th International Congress of Phonetic Sciences, 10–14 August, Glasgow, UK*, 2015.

[27] J. Caspers and V. J. Van Heuven, ”Effects of time pressure on the phonetic realization of the Dutch Accent-lending pitch rise and fall,” *Phonetica,* vol. 50, pp. 161–171, 1993.

[28] A. Arvaniti, D. R. Ladd, and I. Mennen, ”Stability of tonal alignment: the case of Greek prenuclear Accents,” *Journal of Phonetics*, vol. 26, no. 3–25, 1998.

[29] Y. Xu, “Consistency of tone-syllable alignment across different syllable structures and speaking rates,” *Phonetica,* vol. 55, pp. 179–203, 1998.

[30] D. R. Ladd, D. Faulkner, H. Faulkner, and A. Schepman, “Constant ‘Segmental anchoring’ of F0 movements under changes in speech rate,” *J. Acoust. Soc. Am.*, vol. 106, no. 3, pp. 1543–1554, 1999.

[31] M. Atterer and D. R. Ladd, “On the Phonetics and Phonology of ‘Segmental anchoring’ of F0: Evidence from German,” *Journal of Phonetics,* vol. 32, pp. 177–197, 2004.

[32] H. Nam, L. Goldstein, and E. Saltzman, “Self-organization of syllable structure: a coupled oscillator model,” in *Approaches to phonological complexity*, F. Pellegrino, E. Marisco, and I. Chitoran, Eds. Berlin/New York: Mouton de Gruyter, pp. 299–328, 2009.

[33] P. Boersma and D Weenink, “Praat: doing phonetics by computer [Computer program],” version 5.4.01, http://www.praat.org/, 2014.

[34] Y. Xu, “ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis,” in Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France, pp. 7–10, 2013.

[35] R Core Team, “R: A language and environment for statistical computing,” R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/, 2015.

[36] M. Roll, P. Söderström, and M. Horne, “Word-stem tones cue suffixes in the brain,” *Brain Research*, no. 1520, pp. 116–120, 2013.

[37] E. C. Zsiga, ”Acoustic evidence for gestural overlap in consonant sequences,” *Journal of Phonetics*, vol. 22, pp. 121–140, 1994.

Paper II

Research Article

Word-initial consonant-vowel coordination in a lexical pitch-accent language

Malin Svensson Lundmark[1], Johan Frid[2], Gilbert Ambrazaitis[3], Susanne Schötz[4]

[1] Centre for Languages and Literature, Lund University, Lund, Sweden
[2] Lund University Humanities Lab, Lund University, Lund, Sweden
[3] Department of Swedish, Linnæus University, Växjö, Sweden
[4] Logopedics, Phoniatrics and Audiology, Clinical Sciences, Lund University, Lund, Sweden

Short Title: Consonant-vowel coordination in a pitch-accent language

Keywords: articulation; articulatory phonology; coarticulation; EMA; lexical pitch-accent; lexical tone; speech production

Abstract

Previous research has acknowledged the effect of prosody on inter-gestural coordination, but specifically the effect of tones is still understudied. Based on Articulatory Phonology, it has been suggested that tones – in tone languages – are integrated in the gestural organization of a CV-syllable onset in the same manner as would an additional onset consonant (CCV), resulting in a c-center effect, indicated by a CV-onset time lag. However, previous cross-linguistic research has used small-scale data sets and a variety of articulatory measures. This paper presents an investigation using Electromagnetic Articulography (EMA) on word-initial consonant–vowel (CV) coordination in a lexical pitch-accent language (Swedish) with a binary tonal word accent distinction: a tonal fall and a tonal rise, respectively. A selection of 14 spatial or temporal measures of bilabial and tongue body data from 19 speakers, and acoustic f0 data, were examined to study the CV sequence /ma/. Mixed effects regression models revealed differences in the closing of the lips between the two tonal contexts, as well as a longer tongue body movement with a higher positioned trajectory in the rising tone context. Results on CV-synchronization suggest either synchronized CV-onsets or a CV-onset time tag (as in tone languages), depending on the timing landmarks used. Our results partly confirm previous findings, and suggest that these are to some extent explained by methodology (i.e. choice of measures and participants). In addition, our results show that word-initial CV-coordination differs between tonal contexts. We interpret our findings, in the first place, as the result of an assumed synchronization of vocalic and tonal targets and an anticipatory adjustment of the tongue body gesture towards this target. A potential role of the physiological connection between the tongue and the larynx is also discussed. We thus argue that our results rather do not reflect an instance of the c-center effect. However, our results are generally in line with previous results on tone languages in that lexical tone in Swedish has an effect on word-initial CV-(co)articulation.

1 Introduction

This study investigates how the articulation of the word-initial consonant-vowel (CV) sequence /ma/ is affected by the tonal context, using Electromagnetic Articulography (EMA). To this end, the two tonal conditions offered by the pitch-accent language Swedish (Accent 1 and Accent 2) are compared. In Swedish, the binary tonal word-accent distinction is morphophonological, and – in the southern dialect chosen for this study – the tones are executed on the stressed syllable by means of a fall in the fundamental frequency (f0) in Accent 1 and a rising f0 movement in Accent 2.

Previous cross-linguistic articulatory work has aimed to understand how consonant-vowel coarticulation varies as a function of suprasegmental features and phonotactics. For instance, it has been shown that CV sequences exhibit a greater degree of coarticulation, and more phonetic change, than VC sequences (e.g., Fant, 1969; MacNeilage & DeClerk, 1969; Stevens & Blumstein, 1981; Löfqvist & Gracco, 1999; Recasens, 2002), and articulatory coordination varies depending on the types of consonants and vowels involved (e.g., Fowler & Saltzman, 1993; Löfqvist & Gracco, 1999; Recasens, 2002; Mooshammer et al., 2006; Kawahara et al., 2014).

Effects of prosody, both concerning its grouping and its prominence functions, have also been studied within this line of research (e.g., de Jong et al., 1993; Cho, 2002; Byrd et al., 2005; Lindblom et al., 2007; Mooshammer et al., 2013). Recent contributions to the development of this field include work on lexical tones in tone languages, suggesting effects of tone on lip and tongue body coordination (Gao, 2008; Karlin & Tilsen, 2015; Hu, 2016; Zhang et al., 2019), as well as differences in tongue body height between the tones (Erickson et al., 2004; Hoole & Hu, 2004; Shaw et al., 2016).

The few existing relevant studies on the relation between articulatory movements and prosody have so far been based on small-scale data sets consisting of one to seven speakers (e.g., Cho, 2002; Erickson et al., 2004; Hoole & Hu, 2004; Byrd et al., 2005; Lindblom et al., 2007; Gao, 2008; Mooshammer et al., 2013; Niemann et al., 2011; Erickson et al., 2012; Mücke et al., 2012; Karlin & Tilsen, 2015; Hu, 2016; Shaw et al., 2016; Zhang et al., 2019). Small data sets, however, might constitute a serious limitation, since inter-gestural coordination has been shown to be speaker-dependent (e.g., Löfqvist & Gracco, 1999; Gao, 2008; Zhang et al., 2019). In addition, previous studies have used a variety of measures (see 1.5 below), making cross-linguistic comparisons, or comparisons between studies, difficult.

The goal of the present study is therefore to further explore the effect of prosody on inter-gestural coordination by (i) including a considerable number of speakers, (ii) replicating a range of measures that have been used in previous studies, and (iii) extending the field to a previously neglected type of prosodic condition (tones in a pitch accent language).

To this end, we are investigating the effect of the Swedish word accent as a binary categorical distinction on lip and tongue body coordination in the word-initial CV sequence /ma/. The falling and rising tones provide us with two divergent tonal environments, hence two tonal categories. Our analysis is based on kinematic data on bilabial and tongue body movements and acoustic f0 data from 19 speakers. While our main aim is to explore articulatory differences between the two tonal categories, this study is also methodological in that it integrates and compares previously used spatiotemporal measures. This is done in order to better understand the significance of our results by making them comparable to some of the previous work on CV coordination. To the best of our knowledge, the study is the first large-scale articulatory study (in terms of measures and participants) that uses EMA to examine CV coordination in a lexical pitch-accent language.

1.1 Sources of variation in CV coordination

The production of consonants and vowels can be understood as a process in which phonetic gestures are implemented through a coordination of articulators (see Fowler & Saltzmann, 1993, for a more detailed discussion). In the production of a sequence of phonemes, this coordination involves a temporal overlap (or coproduction) of phonetic gestures, which results in anticipatory and carryover coarticulatory effects.

An influential study by Öhman from 1966 revealed how the coarticulation of stop consonants and vowels is the result of overlapping articulatory gestures. In a VCV sequence, the two vowels can be understood as one basic diphthongal gesture, on which the consonantal gesture is superimposed (Öhman, 1966). This overlapping of consonants and vowels has later been shown to differ depending on syllabic position. A CVC structure, for instance, displays more coarticulation in the CV than in the VC sequence (MacNeilage & DeClerk, 1969; Löfqvist & Gracco, 1999; Recasens, 2002). Löfqvist and Gracco (1999) were able to demonstrate that the consonant and the second vowel in a VCV sequence are roughly synchronous with each other, with the vowel gesture onset often even preceding the onset of the consonantal gesture, depending on the vowels involved. Hence, the timing of the gestures in the CV sequence is also sensitive to phonotactic information. Different consonants do indeed exhibit different overlapping patterns with the vowel (Fowler & Saltzman, 1993; Recasens, 2002; Mooshammer et al., 2006), in addition to being dependent on syllable structure (Nittrouer et al., 1988; Byrd, 1996; Byrd et al., 2005). In consonant clusters, timing differences with the vocalic gesture have also been observed between different consonant combinations (Pouplier, 2012; Bombien et al., 2013; Marin, 2013). Finally, tendencies towards language-specific patterns of the CV coordination have also been discussed (e.g., Fowler & Saltzman, 1993; Smith, 1995; Marin, 2013).

1.2 The effect of tone on CV coordination

While the effect of phonotactics on CV coordination is well documented, an area that is still understudied is the effect of tone on the articulators' movements. A major contribution to this line of work can be found within the Articulatory Phonology (AP) framework. The framework proposes articulatory gestures as phonological units (Browman & Goldstein, 1992), and since the pioneering work of Gao (2008) *tone gestures* are assumed as a phonological representation of tone and of the coordinated articulatory actions to achieve the tonal task goal. Tone gestures are comparable to consonantal and vocalic gestures, in that they seem to be integrated in the "global organization" (Browman & Goldstein, 1988) that had previously been observed for consonantal gestures in syllable-initial position: As shown by Gao (2008), the presence of a tone gesture may have a similar effect on the coordination of the CV pattern as the presence of an additional consonant (as in CCV) would have, as detailed in the following paragraphs. Notably, however, such effects have so far not been observed for non-lexical tones (i.e. intonational tones, see Niemann et al., 2011; Mücke et al., 2012).

The assumption of a global organization (as opposed to a local organization) of gestures is based on articulatory evidence from syllable-initial consonant clusters (i.e. CCV- or CCCV-onsets) (Browman & Goldstein, 1988). While in a CV-onset, the consonantal and vocalic gestures are roughly synchronized, for consonants in a cluster (e.g., C_1C_2V) a temporal "shift" of gestural onsets relative to the vocalic gesture onset can be observed, such that C_1 starts slightly before the vocalic gesture and C_2 slightly after (e.g., Browman & Goldstein, 1988; Byrd, 1995; Marin, 2013). That is, as compared to a simple CV syllable onset, C_1 and C_2 in a cluster are "shifted" such that their gestural onsets are timed roughly symmetrically around the vocalic gesture onset. Such findings were interpreted as evidence for a global organization of gestures at syllable onset, and as a metric of this global organization, the c-center (consonant center) was proposed (Browman & Goldstein, 1988; Byrd, 1995). As assumed within AP, it is the c-center that is synchronized with the vocalic gesture in syllable-initial position, rather than the gestural onset of one of the individual consonants in a cluster. As a result of this gestural organization, time lags arise between the onsets of the different articulatory gestures: between C_1 and V, between C_2 and V and between C_1 and C_2. This "c-center effect" (Byrd, 1995) has subsequently been modeled with reference to gestural coupling relationships dependent on syllable position (Nam et al., 2009). According to this account, the onset of the consonantal gesture in a CV sequence is coupled with the onset of the vocalic gesture, and is thus synchronous. When an additional consonantal gesture at onset (as in C_1C_2V) is present, however, the two consonantal gestures are competitively coupled with each other, resulting in the c-center effect and in timing differences between the consonantal and vocalic gesture onsets.

Within this theoretical framework Gao (2008) conducted laboratory work on the four Mandarin tones, measuring the time lag between the gestural onsets of the consonant (C), the vowel (V) and the tone (T). Gao found that the CV time lag between the onsets of the bilabial closing and the tongue body movement in a /ma/ sequence was approximately 45 ms, and likewise approximately 45 ms (30-60 ms for Tone 1, 2, and 3) between the tongue body and the tone gesture onset (VT time lag), where T was measured in the F0 curve (although there was no conclusive result for Tone 4). In other words, the onsets of the consonantal gesture and the tone gesture were timed roughly symmetrically around the onset of the vocalic gesture (C starting 45 ms before, T starting 45 ms after the V onset), which is different from the established pattern attested for single-C syllable onsets (CV) in non-tonal languages, but rather comparable to the attested pattern for consonant clusters (CCV) (e.g., Browman & Goldstein, 1988; Marin, 2013). The time lags were interpreted as the tones being subject to the c-center effect just as consonants are (Gao, 2008). The study was later replicated and tested on Catalan and German pitch accents with no such time lag between C and V onset found (Mücke et al., 2012). Although the CV gestures were synchronous with each other, the onset of the tone gesture was substantially delayed (~100 ms) for the German speaker. A similar study on three German (and two Italian) speakers found similar patterns (Niemann et al., 2011).

The synchronized gestural CV onsets in non-tonal languages (German, Catalan, Italian) have been interpreted by the authors as a result of the absence of lexical tones in these languages (Niemann et al., 2011; Mücke et al., 2012). However, in order to clarify whether the synchronization differences between tonal and non-tonal languages are due to the presence vs. absence of tone, rather than other language-specific parameters, a baseline in the experimental designs would be required, such as a non-tonal condition in the Mandarin data. Zhang et al. (2019) recently made a first attempt to solve this issue by comparing a full tonal condition with a reduced tonal and a non-tonal condition in Mandarin. They found CV time lags in the tonal condition, but no CV time lags in the non-tonal condition. However, they measured gestural targets, rather than onsets. For gestural onsets, they observed, “a high degree of temporal variability” (Zhang et al., 2019).

Moreover, although the proposition of different gestural patterns for tonal and non-tonal languages might be theoretically plausible (CV time lags for lexical tones, no CV time lags for the non-tonal condition), the studies mentioned used different timing landmarks as well as a small number of speakers (seven in Gao, 2008, one from each language in Mücke et al., 2012, and two or three, respectively, for the two languages in Niemann et al., 2011) which might be insufficient considering that there is speaker variability (see, e.g., Löfqvist & Gracco, 1999; Gao, 2008; Zhang et al., 2019). As an intermediate conclusion, we still lack convincing evidence for tone gestures that would affect CV alignment in terms of a c-center effect.

Other articulatory studies on Mandarin tones have observed differences in tongue body movements between the tones. For instance, Shaw et al. (2016), using spatial tongue data, found a lower tongue body in /a/ for the tones with initial low tone context (Tone 2 and Tone 3) than in the initial high tone contexts (Tone 1 and Tone 4). A similar effect of tone was also observed in another study on Mandarin comparing Tone 1 (a high tone) and Tone 3 (a low tone) (Erickson et al., 2004). The tongue was more retracted for Tone 3, and for one of the speakers, also lower. In addition, there was also an effect of tone height on the jaw movement in both of these studies (Erickson et al., 2004; Shaw et al., 2016). As has been shown in a study using Magnetic Resonance Imaging (MRI) by Honda et al. (1999), low tones encompass lower larynx movements, with the jaw moving downwards as well, while higher tones display the jaw moving backward as the high tone falls.

The cited studies on tongue and jaw positions or movement patterns suggest a link between f0 and the articulators, reminiscent of the well documented case of intrinsic f0 in vowels (e.g., Peterson & Barney, 1952; Lehiste & Peterson, 1961; Reinholt Petersen, 1978): low-tongue vowels have a lower intrinsic f0 than high vowels; and low tones seem to trigger a lower jaw/tongue height than high tones. It has been suggested that this connection between F0 and the articulator can be explained by the physiological (muscular) connections between the tongue (the hyoid bone), the larynx (more specifically, the cricothyroid joint) and the jaw (Erickson et al., 2017). The connection between the larynx and the hyoid had already been discussed in connection with the tongue pull hypothesis in the early studies on vowel intrinsic pitch (e.g., Lehiste, 1970; for an overview see Ohala & Eukel, 1987), and a recent study suggests an active role of the jaw, too (Chen et al., 2019).

To conclude, previous studies suggest that the anticipatory movements of CV coarticulation, as well as their spatial and temporal coordination, may, possibly through physiological mechanisms, be affected by the presence of phonological tones, and differently by different tones. The present study extends this line of research to a pitch-accent language, i.e. another type of usage of phonological tone, exemplified by Swedish.

1.3 Tone in Swedish

Swedish has a binary tonal word accent distinction (Accent 1, Accent 2, henceforth A1 and A2) with tonal phonological representations and phonetic realizations that differ largely between dialects (Bruce & Gårding, 1978). In this study, we examine Southern Swedish, where A1 is represented by an initial high tone at vowel onset (a falling tone, H*L, in the stressed syllable), and A2 by an initial low tone at vowel onset (a rising tone, L*H) (Fig. 1). Swedish word accent is morphophonological: Words with final lexical stress (which includes all monosyllabic words) always bear A1, while in polysyllabic words with non-final stress, both accents can appear. In these words, they can have either lexical or morphological uses, as they are, for instance, associated with different suffixes. Phonologically, however, the falling and the rising tone are both associated with the stressed syllable (located in the stem, not in the suffix), which results in an early f0 peak in A1, and a late f0 peak in A2. Previous studies have shown that the early and the late timing of the f0 peaks are consistent, and perceptually important (Bruce & Gårding, 1978; Bruce, 2005; Ambrazaitis & Bruce, 2006; Ambrazaitis et al., 2012; Svensson, 2014).

Details of the co-production of the Swedish tones and the CV sequence have been hardly addressed in the past, and the few exceptions have been based on acoustic data. Early studies by Öhman (1965) and Löfqvist (1975), have, for instance, studied F0 contours in varying segmental conditions aiming to establish invariant F0 properties of the word accents (Öhman, 1965), or, in other words, to disentangle intrinsic from extrinsic F0 variations (Löfqvist, 1975). Other acoustic studies have revealed durational effects of the Swedish word accents (Elert, 1964; Svensson Lundmark et al., 2017). Another more recent example is a study on tonal-segmental alignment in South Swedish (Svensson, 2014). Articulatory measurements in connection with the Swedish word accents have, to the best of our knowledge, been restricted to investigations of laryngeal control (e.g., Öhman et al., 1967; Gårding et al., 1975). Little is known about whether or how the articulatory movements of consonants and vowels in Swedish are affected by the tones of the word accents.

1.4 Scope and limitations.

Previous articulatory studies on tones suggest that tones are integrated with the articulation of the consonant and the vowels in a CV syllable onset (see 1.2). This integration is evidenced by time lags between the articulatory onsets of C and V in tonal languages, while previous results on non-tonal languages suggest a synchronicity of the C and V onsets (although these results are not entirely conclusive, as discussed in 1.2). Within Articulatory Phonology, such findings have been interpreted as evidence of a tone gesture that is coupled with the consonantal and the vocalic gesture in much the same way as the two consonantal gestures and the vocalic gesture are coupled in a CCV cluster. Since the lexical pitch accents of Swedish are understood as consisting of tones, we would expect that their hypothetical tone gestures are coupled with the consonantal and vocalic gesture in a CV syllable onset, just as in a tone language. Hence, for the present study, we would predict time lags between the consonantal and the vocalic gesture onsets in a CV syllable onset. Furthermore, we might expect to see differences in tongue body movements between A1 and A2, in line with the findings for Mandarin tones by Erickson et al. (2004), Hoole and Hu (2004), and Shaw et al. (2016).

Therefore, in this study we focus on the effect of tonal context on the lips and the tongue body and present spatio-temporal data on 1) the bilabial closing and release during the word-initial consonant, 2) the tongue body during the lowering and retraction of the simultaneous vowel movement, and 3) the time lags between these separate articulator movements. To this end, we compare the measures just mentioned in two conditions: /ma/ word onsets in words with A1 vs. words with A2.

The scope of this study is to take this first step of testing whether CV-coordination is affected by the tone (T). A subsequent step, not taken in this study, could then be to establish a measure of the onset of a tonal gesture and to include it in the study of temporal coordination, measuring time lags between C, V, and T.

Initial attempts of including T have been made in the studies by Gao (2008) and Mücke et al. (2012) (but see the discussion in 4.4 below). The present material, however, would not allow us to present results on V and T time lags in comparison to the studies by Gao (2008) and Mücke et al. (2012). The initial low tone gesture of A2 presupposes a preceding fall from a high tone, a constellation that is not possible to produce in South Swedish without inserting a phrase boundary. Although it would be possible to collect data on and compare the tonal *targets* (instead of onsets) of A1 and A2, within the scope of the present study, there is no theoretically grounded motivation to do so.

Therefore, instead of measuring tonal alignment as has been done in Gao (2008) and Mu¨cke et al. (2012), we undertake a different approach to the f0 data. As already mentioned, we treat the word accents as two different tonal categories (as has been similarly done on Mandarin in Shaw et al., 2016) and present a visual representation of the overall f0 pattern as well as an analysis of f0 differences between the word accents in the most relevant region of interest, that is during the word-initial consonant /m/.

1.5 Methodological considerations

Previous related studies cited above have made use of a variety of methodological choices, which are not always compatible and to some extent inhibit direct comparisons. In the present study, we aim to avoid this problem and have hence opted to include a variety of alternative measures, enabling such comparisons. In this section, we motivate our choice of measures, rather than explaining them in detail (although we cannot avoid a certain degree of detail here). However, we recommend consulting Figures 2 and 3 below for visual support in following the present discussion, if needed.

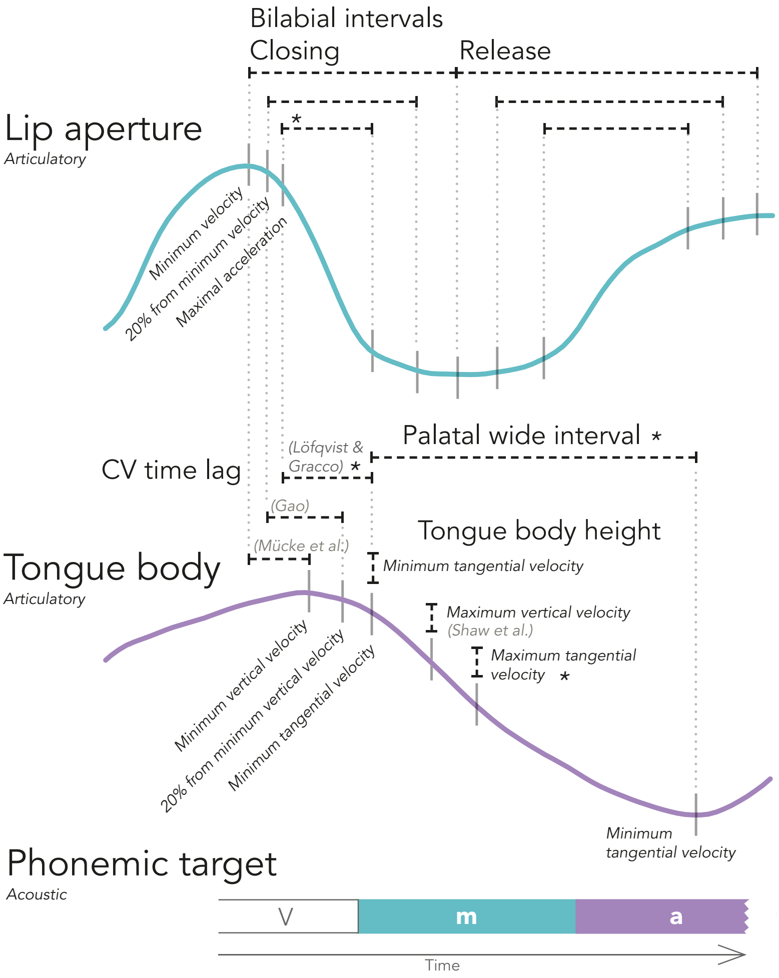


Figure 2. An illustration of the spatiotemporal measurements on the bilabial (Lip aperture) and tongue body data of the target CV sequence /ma/. The figure is a stylized visualization of the approximate position of each measurement. Some of the measurements occur already during the previous vowel. (* significant differences between the word accents found after the analysis). See Fig. 3 for more detail.

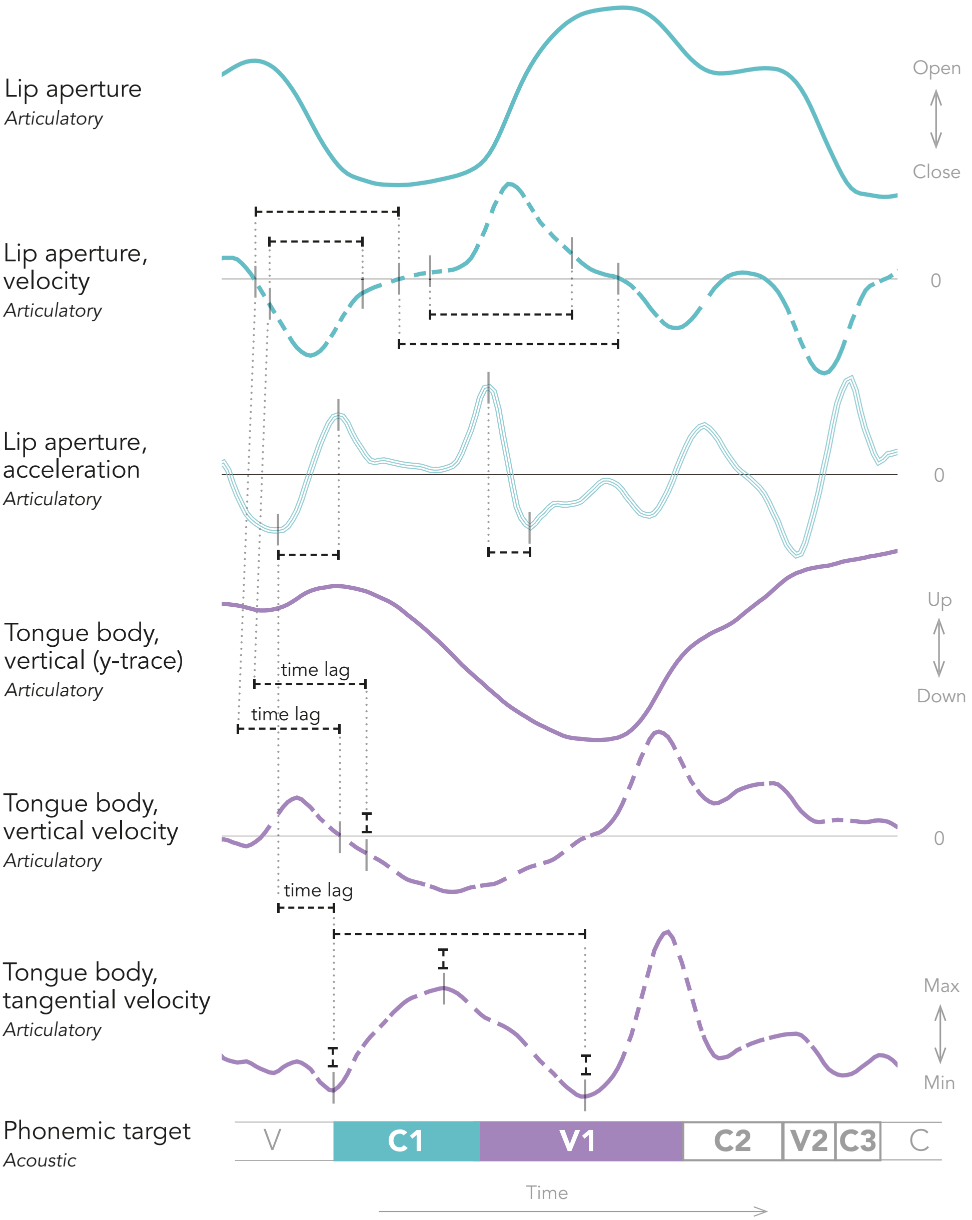


Figure 3. A detailed visualization of the spatiotemporal measurements. Lip landmarks are based on the first derivative (velocity) and second derivative (acceleration) of the Lip aperture calculation. Tongue body landmarks are based on the first derivative of the actual TB sensor movement: both y-trace (vertical velocity) and on the combined x and y (tangential velocity).

The EMA data was obtained from speakers pronouncing A1 and A2 words with onsets containing a nasal bilabial consonant and an open vowel (/ma/). Most previous articulatory studies have included this specific CV sequence, which makes it useful for comparisons. For example, the dataset in the dissertation by Gao (2008) included bilabial and alveolar nasal consonantal onsets, as well as voiceless onsets, followed by an open vowel. Mücke et al. (2012), in their study of Catalan and German, compared bilabial and alveolar onsets followed by open vowels, in both open and closed syllables.

In order to measure gestural onset-to-onset (CV) time lags, both Gao (2008) and Mücke et al. (2012) based their measurements on the vertical movements of the tongue body (for V) and on the calculated lip aperture (distance between lip sensors) for C. However, while Gao (2008) labelled the C and V landmarks on the velocity curve at a 20% threshold from the zero-crossing to peak velocity (i.e. to the local minimum in Fig. 3), in Mücke et al. (2012) the corresponding landmarks were instead placed at the zero-crossings (also on the velocity curve), that is, when the vertical movement was minimal. It is unclear how or if the type of landmark affects the duration of the time lags. However, a threshold mark of 20% from zero crossing velocity might be preferable, since a plateau may occur around the zero-crossing in the velocity curve, creating several possible landmarks. In addition, in previous studies, the perhaps more stable 20% threshold mark has been found to be advantageous for capturing an on-going movement (Kroos et al., 1997; Mooshammer et al., 2006). We replicated and compared these two measurements in our study of Swedish tones.

Even though Löfqvist and Gracco (1999) did not include tone in their study on the coarticulation of VCV sequences, the methodology is of potential interest for our study. They captured the dynamic tongue body movements using data on both tongue height and tongue frontness (horizontal and vertical dimension) by calculating the tangential velocity. Tangential velocity, [$v = \sqrt{(\dot{x}^2 + \dot{y}^2)}$], captures the dynamic tongue body movements and combines the front-back and the high-low dimensions (Löfqvist & Gracco, 1999). Furthermore, since they investigated several vowel types, they found that a lip aperture landmark based on acceleration, that is, the second derivative of the lip movement, was the most stable across the vowels. We believe that their methodology could add to the comparison with Gao (2008) and Mücke et al. (2012) since it will provide us with a wider account of the dynamics of the lips and tongue body movement.

Shaw et al. (2016) examined spatial dimensions of a set of three different vowels following a bilabial consonant and found specifically for the syllable /pa/ that tongue body and tongue tip sensors differed in height between the Mandarin tones. They measured at a landmark in the velocity curve for the tongue sensors, at a 20% threshold from the zero-crossing to peak velocity, because they found it to be a stable anchor point for measuring tongue body height (Shaw et al., 2016). Incidentally, this is the same landmark on the tongue body used in the study by Gao (2008), where it represented the onset of the tongue body movement. Erickson et al. (2004), in their study on Tone 1 and Tone 3, instead used the vertical and the horizontal position (measured separately) of a sensor on the tongue dorsum at a time when the jaw was at its lowest point. Their landmarks therefore reflect the target of the tongue body gesture. In order to be able to compare our spatial tongue body results on the Swedish word accents to both of these studies (Erickson et al., 2004; Shaw et al., 2016), we include (1) the measure of tongue height at the threshold of peak velocity of the tongue (Shaw et al., 2016), as well as (2) a measure that aims to capture tongue height at the vowel target, thus resembling the measures by Erickson et al. (2004). However, we base our measure of tongue height at vowel target on the tangential velocity curve. This way, we combine the two measures used by Erickson et al. (2004), who saw an effect of tone on both the vertical and horizontal position of the tongue dorsum. Using the tangential velocity curve is also motivated by Löfqvist and Gracco (1999), who used it to measure the CV onset time lag.

In addition, to gain a more comprehensive picture of the tongue movement, we measure tongue height at the following landmarks in the tangential velocity curve: minimum tangential velocity at the onset (used to measure the CV onset time lag, here and in Löfqvist and Gracco, 1999), minimum tangential velocity at the target (which reflects Erickson et al., 2004) and maximum tangential velocity. The latter landmark we found as a stable point to reflect a point in time between the onset of the movement and the time of reaching the target, that is, during the lowering of the tongue.

Since we have previously noticed the tongue body lowering to differ between A1 and A2 (Svensson Lundmark et al., 2017), we include a measurement on the duration of the tongue body movement: the *palatal wide interval*. This term refers to the time it takes for the tongue body to reach its target (the lowest position).

We also added bilabial interval measurements of the lips in order to evaluate the CV time lag measurements (by “CV time lag” we refer to the difference between the onset of the bilabial closure and the onset of the tongue body movement). Accordingly, "bilabial interval" refers to measured intervals for the time it takes for the lips to close and to open again. Both intervals are measured in three different ways: Since we are including three different landmarks to measure the bilabial closure onset, based on previous studies (see above), we also use the same landmarks to measure closure offset, and for the sake of consistency, we use the same three types of landmarks for the bilabial release interval. The bilabial intervals are thus either based on when the movement of the lips is minimal (zero-crossing in the velocity curve), based on when the movement is on the way or has slowed down (the 20% peak velocity threshold), and based on when the lips accelerate and decelerate most (maxima and minima in the acceleration curve).

These considerations result in a comprehensive selection of measures to be included in the present study. An overview is presented in Table 1.

Table 1. The different measures and their calculations based on the assortment of articulatory landmarks. ^{1} = the local *minimum* in the acceleration curve (see Footnote 1 for an explanation).

Measurement	**Calculation**	**Lip aperture landmark**	**Tongue body landmark**	**Dependent variable**
Bilabial closure interval	LA target – LA onset	zero velocity	-	C_closure_0vel
		20% zero velocity	-	C_closure_20vel
		max. acceleration/ deceleration	-	C_closure_acc
Bilabial release interval	LA offset – LA target	zero velocity	-	C_release_0vel
		20% zero velocity	-	C_release_20vel
		max. acceleration/ deceleration	-	C_release_acc
Palatal wide interval	TB target – TB onset	-	min. tangential velocity	V_interval
Tongue body height	y-score at landmark	-	20% min. vertical velocity	V_height_Shawetal
		-	min. tangential velocity	V_height_onset
		-	max. tangential velocity	V_height_peak
		-	min. tangential velocity	V_height_target
CV time lag	TB onset – LA onset	zero velocity	zero vertical velocity	CV_timelag_Mückeetal
		20% zero velocity	20% zero vertical velocity	CV_timelag_Gao
		max.^{1} acceleration	min. tangential velocity	CV_timelag_L&G

2 Materials and Methods

2.1 Speakers

Twenty speakers (12 female) of the South Swedish dialect of Scania (the southernmost region of Sweden) participated in the EMA recordings. The speakers' ages ranged from 23 to 75 (average 40 years, standard deviation 12.3 years). The participants were unaware of the purpose of the study. All speakers grew up with South Swedish speaking parents. However, four of the speakers (20%) grew up with one parent with a first language other than Swedish, and one speaker (5%) with one parent with another Swedish dialect.

2.2 Speech material

The original data set consisted of 18 target words, divided into nine word accent pairs with identical stress pattern, that is, stress on the initial syllable (e.g., *bilen-bilar* or *manen-manar*). For this study, we only used the four word accent pairs that shared the similar word-initial CV sequence /ma/. Hence, the material consists of eight disyllabic target words, four with A1 and four with A2, with phonologically similar word onsets, namely either the target CV sequence /ma/ ([ma]) or /maː/ ([mɑː]). Both comprise a nasal bilabial closure and a palatal wide tongue body movement, but the vowels differ in quality and quantity. Moreover, in the case of the short vowel, the target CV sequence is followed by a coda: either a nasal alveolar, /ˈCVn.nV(C)/, or a nasal bilabial, /ˈCVm.mV(C)/. If the vowel is long it is followed by a nasal alveolar, /ˈCVː.nVC/, or a lateral alveolar, /ˈCVː.lVC/, in post-syllabic position (Tab. 2). Hence, the short condition implies a closed syllable, while the long condition signifies an open syllable.

The eight disyllabic target words were embedded in individual but similarly structured target sentences. The VCV sequence preceding the target words was identical across all sentences (/ade/). Each target sentence was preceded by a leading question eliciting a contrastive focus on the last element in the target sentence (Tab. 2). This left the target word in a low-prominence inducing context, ensuring it would not be associated with a sentence-level prominence. In Central Standard Swedish, sentence-level prominence is typically associated with an additional f0 peak (Bruce, 1977). This kind of categorical distinction between higher and lower prominence is absent from the South Swedish dialect (Bruce & Gårding, 1978; Bruce, 2005). However, higher-level prominence has nevertheless been shown to gradually affect the accentual f0-pattern even in South Swedish (Ambrazaitis et al., 2012) and, under certain pragmatic conditions, a double-peaked pattern can be observed in this dialect, too (Svensson Lundmark et al., 2017).

Therefore, effects of sentence intonation were controlled for in this study by means of the low-prominence inducing context.

Table 2. Stimuli material arranged according to the conditions *vowel length* and *word accent* (A1, A2). Target words in SMALL CAPS; note that main sentence prominence is elicited on the final word.

		Leading question and target sentence
Short vowel (V)	A1	”Hur rammade Eva mannen? Eva rammade MANNEN med cykeln.” *How did Eva hit the man?* *Eva hit the man with her bike.*
	A2	”Hur skummade Inger manna? Inger skummade MANNA med vispen.” *How did Inger foam semolina?* *Inger foamed semolina with the whip.*
	A1	”Med vem kollade Sanna filmen ’Mammut’? Sanna kollade ’MAMMUT’ med Nora.” *With whom did Sanna watch the movie 'Mammoth'?* *Sanna watched 'Mammoth' with Nora.*
	A2	”Var lämnade morfar mamma? Morfar lämnade MAMMA med doktorn.” *Where did grandpa leave mom?* *Grandpa left mom with the doctor.*
Long vowel (V:)	A1	”Vilken häst kammade Jonas manen på? Jonas kammade MANEN på Ronja.” *Which horse did Jonas comb the mane of?* *Jonas combed the mane of Ronja.*
	A2	”Vad rimmade Isak ’manar’ med? Isak rimmade ’MANAR’ med ’svanar’.” *What did Isak rhyme ‘urge’ with?* *Isak rhymed 'urge' with 'swans'.*
	A1	”Var lämnade Inger malen hon fångat? Inger lämnade MALEN till kocken.” *Where did Inger leave the sheatfish she caught?* *Inger left the sheatfish to the chef.*
	A2	”Var lämnade Åsa alla malar hon fångat? Åsa lämnade MALAR vid båten. *Where did Åsa leave all the sheatfish she caught?* *Åsa left sheatfishes by the boat.*

2.3 Recording procedure and data preprocessing

Speakers were recorded with the Carstens AG501 Electromagnetic Articulograph (EMA). Data was collected from sensors on the midline of the upper and lower lips, at the vermillion border, and from one sensor placed on the midline of the tongue body. The tongue body sensor, corresponding roughly to the tongue dorsum, was placed on the tongue where the participant made a mark with the upper incisors after having stretched out the tongue as far as possible. The movement of the jaw is integrated with the data on the lips and the tongue and was not treated separately, although we do acknowledge the effect of the jaw on the lip aperture and on the tongue body movements. In order to correct for head movements, we used three additional reference sensors: one behind each ear, and one on the nose ridge. The occlusion plane was not controlled for other than by observing and correcting the speaker's position during the recording. However, the head rotation angle was not expected to have an effect on the combined x and y dimension, or on the distance measurements. While the vertical tongue body measurements could be affected by the head rotation angle, this was handled in terms of speaker variation in the statistical analysis, where we assumed random intercepts for subjects in a mixed effects model, as well as random slopes when warranted.

The eight sentence pairs (leading + target sentences) were read from a computer screen in a random order, each set appearing eight times. Articulatory data was recorded at 250 Hz, and audio was recorded simultaneously using an external condenser microphone (a t.bone EM 9600) at a sampling rate of 48 kHz.

Before subjected to analysis, the data were corrected for head movements with the Carstens software, using the three reference sensors mentioned above, and transferred to R (R Core Team, 2015). The articulatory data was smoothed using locally weighted regression by the R function *loess*. We used a low span (0.1) so the smoothing should not cause any distortion. This value was determined by testing different span values on a subset of the data and visually inspecting the result.

The first author segmented the acoustic data manually in PRAAT (Boersma & Weenink, 2018) into consonants and vowels (C_1, V_1, C_2, V_2, C_3) using the PRAAT script ProsodyPro (Xu, 2013). In order to avoid unnecessary f0 analysis errors, f0 calculation was performed in the time-domain based on “pulses” automatically determined by PRAAT, which we manually corrected using ProsodyPro. The script also applies a smoothing algorithm removing minor spikes from f0 curves (Xu, 1999). For the visualization of the tonal patterns of the two word accents (Fig. 1), f0 was time normalized by taking ten temporally equidistant f0 measurements for each segment. Finally, these corrected, trimmed, and time normalized f0 data (in Hz) were converted into semitones (using 100 Hz as a reference value). For the purpose of visual inspection, these pre-processed f0 data were then used to plot mean f0 curves across repetitions and speakers, but separated for the eight target words (Fig. 1).

For the statistical analysis of the f0 differences during the word-initial /m/, the individual words were time-normalized by z-transforming the time points (i.e. subtracting the mean and dividing by the standard deviation; the mean and standard deviation were obtained on the basis of all time points in the word). Furthermore, f0 measurements were normalized for each word by z-transforming the measurements per word on the basis of all measurements in the utterance it occurred in.

All speakers but one (who had one parent with another Swedish dialect), displayed the typical South Swedish f0 pattern of the word accents, with an f0 fall in the stressed vowel for A1, and a rise in the stressed vowel for A2 (Fig. 1). The speaker with the deviant f0 pattern was excluded from subsequent analysis. With repetitions, and some omissions due to interferences in the signal during the recordings, the data collection ended up consisting of 1191 tokens.

2.4 Measurements

The PRAAT TextGrids were further used in R (R Core Team, 2015) to automatically extract articulatory data on lip and tongue body movements from selected time frames of each target word. These time frames were based on acoustic segment boundaries and adjusted manually for each speaker. A constant time frame for all speakers (starting, say, 30 ms before the acoustic segmental consonant onset) would be problematic, because inter-gestural coordination and speech rate vary between speakers. After a visual inspection of the position, velocity and acceleration curves, the first author adjusted the time frames in the script for each individual speaker.

Lip aperture (LA) was calculated in R as the three-dimensional Euclidian distance between the sensors on the upper and the lower lip. For the bilabial intervals temporal landmarks on lip aperture were automatically extracted in R from when the lips had minimal movement (zero-crossing in the first derivative, i.e. velocity curve, see Fig 3.), when the movement was on its way or when it had slowed down (20% threshold from zero to peak velocity, i.e. to local minimum/maximum in the first derivative, i.e. velocity curve), and when the movement accelerated or decelerated the most (local minima/maxima in the second derivative, i.e. acceleration curve). First and second derivative were calculated by the R diff function with lag=2 (R Core Team, 2015).

The landmarks at the onset of the movement towards the bilabial closure (C) and at the onset of the lowering (and the retraction) of the tongue (V) were further used to calculate CV time lags. We have replicated the measurements of CV time lags from Mücke et al. (2012), which are based on the landmarks: zero-crossing in the velocity curve for lip aperture and the vertical tongue body movement; from Gao (2008), based on the 20% threshold from zero-crossing to peak velocity for lip aperture and the vertical tongue body movement; and from Löfqvist and Gracco (1999), based on maximal acceleration (the local minimum in the second derivative, i.e. the acceleration curve)[1] of lip aperture and the minimum tangential velocity of the tongue body (Fig. 2 and 3).

The tangential and the vertical velocity of the tongue body were further used to also measure the tongue body height. Similar to Shaw et al. (2016), spatial data was obtained from the same landmark as in Gao (2008): the 20% threshold to peak velocity of the vertical tongue body movement. Moreover, to capture both the lowering and the retraction of the tongue, as has been reported in Erickson et al. (2004), landmarks were placed at minimum tangential velocity (onset and target) and at peak tangential velocity of the tongue body movement (Fig. 2 and 3). The collected landmarks at the onset and the target of the tongue body movement were also used to extract temporal data in order to calculate the palatal wide interval. Table 1 shows the measurements and calculations resulting in the different dependent variables.

2.5 Analysis

First, we tested whether f0 indeed differed between the word accents during the word-initial consonant by performing a Welch Two Sample t-test on the f0 data. Secondly, we used linear mixed effects regression models to investigate the effect of the main predictor *word accent* on each of the dependent variables (the articulatory measurements). A mixed effects model is justified since both individual speakers' articulation and sensors' placement on the articulators might vary. Hence, the random effect *speaker* (random intercepts) was included in all models fitted on the measurements (random slopes for speakers were only added when the complexity was warranted, as explained further down in this paragraph). Moreover, a qualitative f0 analysis revealed differences between target words (see the f0 analysis in 3.1). Some of the target words might have attracted more prominence by some speakers, possibly because the information structure of the corresponding target sentences enabled inserting phrase boundaries or in other ways emphasizing the target words. Consequently, we added *word* as random effect (random intercept) to avoid target sentence information structure or other factors (e.g., word frequency) increasing the risk of reporting p-values that are too low (a Type I error). Furthermore, although our main interest was the effect of word accent, the covariate *vowel length* (which in our material corresponds to syllable type), and its interaction with word accent, needed to be controlled for. We performed a likelihood ratios test for each measurement to test whether the additional complexity of vowel length was warranted. The complexity was

[1] Note that we refer to this point in time as the maximal acceleration (of the lip closing movement), despite the fact that it corresponds to a local minimum in the acceleration curve. The polarity of the extremum is negative, because the lip aperture curve goes down.

only added when the model comparison resulting in the more complicated model showed a lower Akaike Information Criterion (AIC) value of at least 2 (following Wieling & Tiede, 2017). Additional complexity that was tested for in this way was random slopes by *speakers* as random effect to control for speaker variability. Hence, the mixed effects regression model included the predictor *word accent* as fixed effect, with the random effects *speaker* (random intercepts) and *word* (random intercepts), with the optionally added complexity of the fixed effect *vowel length* and/or the random slopes of the random effect *speaker.*

Outliers were excluded by the use of time frames during the automatic data collection in R. In case of negative values for the intervals (the bilabial intervals and the palatal wide interval), these were replaced by missing values and excluded from the regression analysis. Visual inspection of residual plots did not reveal any obvious deviations from homoscedasticity or normality. The mixed models were run in R using the lme4-package (Bates et al., 2015). P-values were obtained by using the lmerTest-package in R (Kuznetsova et al., 2017).

3 Results

3.1 Acoustic results

A qualitative inspection of the mean f0 contours (Fig. 1) shows that A1 starts with a high tone at vowel onset, followed by a fall during the vowel, while A2 starts with a low tone at vowel onset, followed by a rise during the vowel and an f0 peak at around the syllable boundary. The absolute average f0 peak can be seen to differ in height between the target words, which is most clearly seen in Figure 1(b), where f0 reaches, on average, slightly higher values in *manen* and *manar*. This might at least partially be related to the observation that individual target words have attracted a higher level of prominence (see 2.5 above).

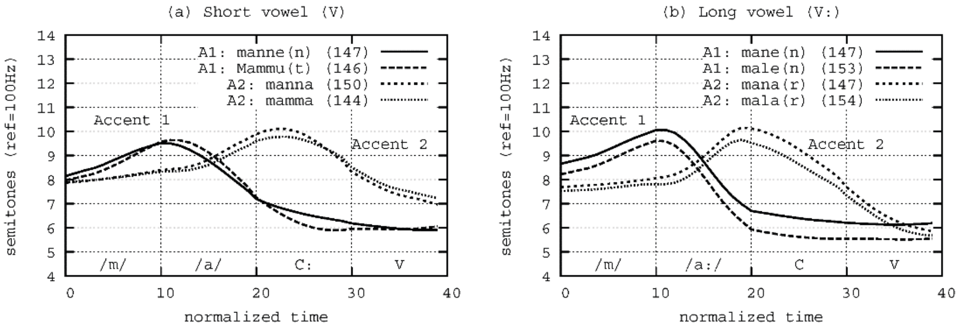


Figure 1. Mean f0 contours of South Swedish speakers in Accent 1 and Accent 2 target words: short vowels (a); long vowels (b). The normalized time scale indicates the points of measurement (ten temporally equidistant measurements per segment); the vertical lines indicate segment boundaries.

Next, a statistical analysis was performed to find out whether f0 differed between the word accents already during the word-initial C_1 segment. The f0 mean of the first segment of the word accents, /m/, differed significantly according to Welch's *t*-test, *t*(12877) = 113.08, *p* < .001. On average, mean f0 for A1 was 1.16 semitones higher than for A2 (Fig. 4). The 95% confidence interval for the effect of word accent on mean f0 was between 1.14 and 1.18 semitones. In other words, although the falling and the rising tone of the Swedish word accents are executed mostly during the vowel segment, there is a significant difference between the f0 environments of the word accents already when the tongue body is lowering, that is during the span that includes the majority of our articulatory measurements (see Fig. 2 for reference).

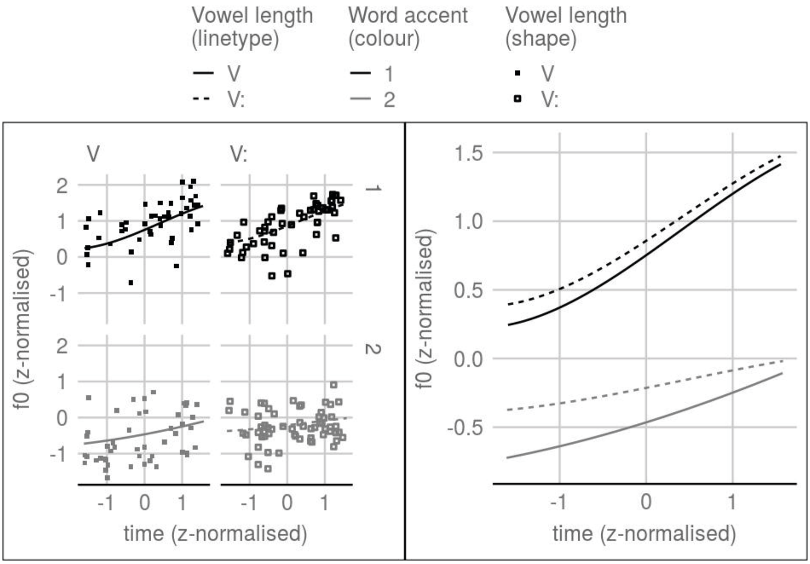


Figure 4. f0 in the first consonant, grouped by vowel length and accent. In each subplot to the left, a line shows a polynomial fit through all data points in each subgroup. The curves around the line show a 2D kernel density estimation. For esthetic reasons, only a random subsample of the f0 points are shown (left). The plot to the right shows the polynomial regression lines of the normalized f0 data during the word-initial consonant, grouped by word accent and vowel length.

3.2 Articulatory results

3.2.1 Bilabial intervals

3.2.1.1 Bilabial closure

After a model comparison for the bilabial closure measurements, a mixed effects regression model with only *word accent* as fixed effect, and *speaker* and *word* as random effects (with random intercepts) was fit on the basis of the temporal landmarks from when the lip movement was minimal (C_closure_0vel) and from when the movement was on its way or was slowing down (C_closure_20vel). The models did not show significant differences between the word accents (Tab. 3). In contrast, for the landmarks at when the lip movement accelerated/decelerated (C_closure_acc), a more complex model was warranted with the added fixed effect *vowel length*, which showed a significant difference between the word accents (t = -2.1, p < .05), as well as between the vowel lengths (t = -3.0, p < .01): A1 and the long vowels displayed slightly longer intervals than A2 and the short vowels (see Tab. 3). Figure 5a visualizes the small but statistically significant differences for the de-/acceleration landmarks (C_closure_acc) compared to the non-significant differences with velocity landmarks (middle and left panel in Fig. 5a). The figure also displays the fact that the intervals calculated from when the lips de-/accelerated (C_closure_acc) and when the movement was on its way/slowed down (C_closure_20vel) are much shorter but also less varied than the intervals from when the lip movement is minimal (C_closure_0vel).

3.2.1.2 Bilabial release

The bilabial release intervals differ in some regards from the bilabial closure. First, as opposed to the closure, the release intervals calculated from the velocity landmarks (C_release_0vel and C_release_20vel) warranted added complexity to the mixed effects regression models. The models with the added *vowel length* as fixed effect showed a clear significant difference between the vowels (C_release_0vel: t = -5.5, p < .001; C_release_20vel: t = -6.3 p < .001): Long vowels displayed longer intervals than short vowels (see Tab. 3 and Fig. 5b, C_release_0vel and C_release_20vel). Second, the model fitted for the de/-acceleration-based interval (C_release_acc) warranted adding random slopes by speaker to the model, but showed no significant difference between the word accents, in contrast with the result on the bilabial closure interval. Note that neither of the bilabial release measurements showed a significant difference between the word accents, although the panels in Figure 5b suggest a tendency towards longer release intervals for A2.

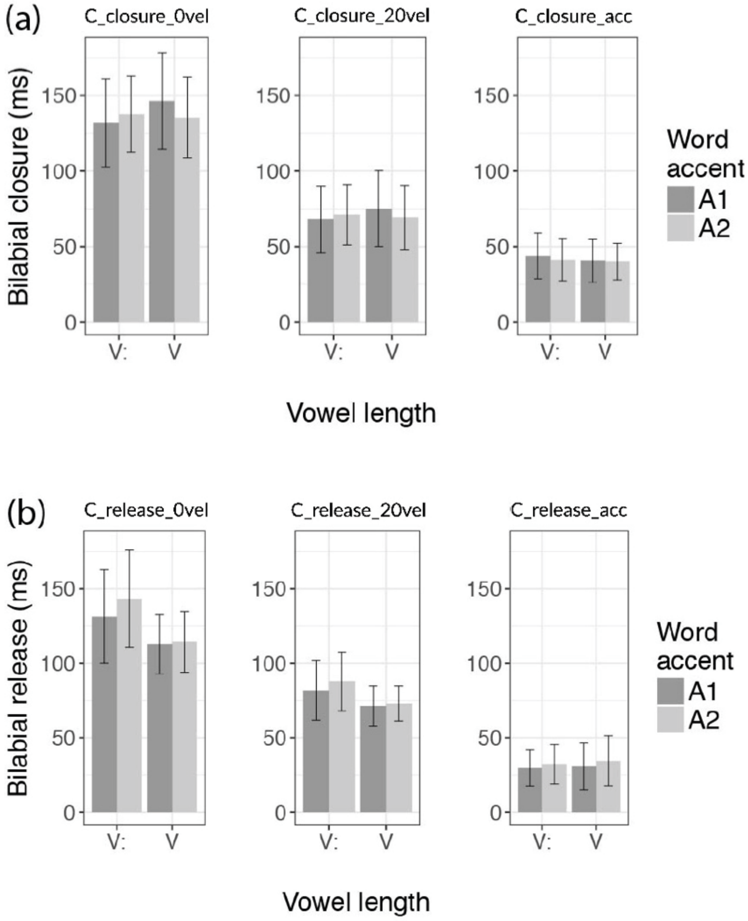


Figure 5. Bilabial closure (a) and bilabial release (b) as an effect of word accent and vowel length: when the lip movement is minimal (C_closure_0vel and C_release_0vel), when the movement is on its way or slowing down (C_closure_20vel, and C_release_20vel), and when the lips de-/accelerate (C_closure_acc and C_release_acc). Error bars represent standard deviations.

Table 3. Results from the mixed effects regression models, with estimates on the effect by word accent and by the addition of interaction with vowel length. * Significant effects. ^{R} =Random slopes by speaker added to the model. ^{AIC} = the more complex model had a less AIC value of 1.4 (as opposed to 2).

Measurement	**Dependent variable**		**Estimate**	**SE**	**df**	**t-value**	**p-value**
Bilabial closure interval (ms)	C_closure_0vel	Intercept	135.09	6.615	24.162	20.409	0.000
		Word accent (A1)	-3.254	8.404	7.639	-0.387	0.709
	C_closure_20vel	Intercept	68.601	4.440	24.322	15.452	0.000
		Word accent (A1)	-1.913	5.574	7.629	-0343	0.741
	C_closure_acc	Intercept	41.460	1.2033	19.000	34.470	0.000
		Word accent (A1)	-1.605	0.751	1170.100	-2.138	0.033*
		Vowel length (V:)	-2.227	0.751	1170.300	-2.966	0.003*
		Interaction	2.366	1.501	1170.100	1.576	0.115
Bilabial release interval (ms)	C_release_0vel	Intercept	124.891	4.146	24.636	30.121	0.000
		Word accent (A1)	6.478	4.462	7.424	1.452	0.187
		Vowel length (V:)	-24.747	4.463	7.435	-5.545	0.001*
		Interaction	-11.389	8.923	7.424	-1.276	0.240
	C_release_20vel	Intercept	78.321	2.317	23.480	33.805	0.000
		Word accent (A1)	3.782	2.054	7.337	1.842	0.106
		Vowel length (V:)	-12.854	2.055	7.357	-6.255	0.000*
		Interaction	-4.731	4.107	7.337	-1.152	0.285
	C_release_acc ^{R}	Intercept	31.801	1.889	23.669	16.840	0.000
		Word accent (A1)	3.214	2.329	8.414	1.380	0.203
Palatal wide interval (ms)	V_interval	Intercept	194.685	5.034	23.833	38.672	0.000
		Word accent (A1)	13.770	5.004	7.459	2.752	0.027*
		Vowel length (V:)	-35.525	5.004	7.462	-7.099	0.000*
		Interaction	-5.884	10.008	7.460	-0.588	0.574
Tongue body height (mm)	V_height_Shawetal	Intercept	10.945	4.311	19.026	2.539	0.020
		Word accent (A1)	-0.076	0.240	6.007	-0.316	0.763
	V_height_onset	Intercept	11.218	4.336	19.037	2.587	0.018
		Word accent (A1)	0.081	0.284	7.006	0.285	0.784
	V_height_peak	Intercept	8.890	4.221	19.000	2.106	0.049
		Word accent (A1)	-0.231	0.098	1043.000	-2.364	0.018*
		Vowel length (V:)	-0.246	0.098	1043.000	-2.515	0.012*
		Interaction	0.973	0.195	1043.000	4.979	0.000*
	V_height_target ^{AIC}	Intercept	6.477	4.175	19.003	1.551	0.137
		Word accent (A1)	-0.344	0.134	5.904	-2.569	0.043*
		Vowel length (V:)	-0.115	0.134	5.900	-0.859	0.424
		Interaction	0.796	0.268	5.912	2.972	0.025*
CV time lag (ms)	CV_timelag_Mücketal	Intercept	60.237	8.613	24.894	6.993	0.000
		Word accent (A1)	-6.367	10.404	7.593	-0.612	0.558
	CV_timelag_Gao	Intercept	67.769	7.455	25.486	9.090	0.000
		Word accent (A1)	-4.807	8.358	7.527	-0.575	0.582
	CV_timelag_L&G	Intercept	1.422	3.081	19.000	0.462	0.650
		Word accent (A1)	-8.157	1.688	1140.200	-4.830	0.000*
		Vowel length (V:)	6.211	1.690	1140.400	3.676	0.000*
		Interaction	-0.497	3.377	1140.200	-0.147	0.883

3.2.2 Palatal wide interval

For the calculated intervals of the tongue body lowering and retraction (palatal wide interval, Fig. 6), the model with random slopes did not differ significantly from the model without random slopes. Instead, the model comparison warranted the added complexity of *vowel length* as fixed effect. The model fit on the palatal wide interval (V_interval) showed a significant difference between word accents (t = 2.8, p < .05), as well as between vowel lengths (t = -7.1, p < .001) (see Tab. 3). The estimated average duration was 195 ms and the A2 intervals were approximately 14 ms longer than the A1 intervals (in the short vowel condition). Moreover, the interval was 36 ms longer for long vowels than for short vowels (in the A1 condition). There was no significant interaction, and thus, the effect by word accent on the palatal wide interval is found to be independent of vowel length.

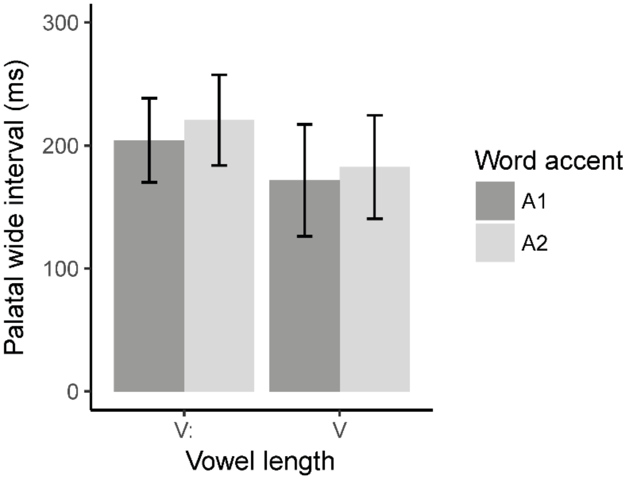


Figure 6. Palatal wide interval as an effect of word accent and vowel length.

3.2.3 Tongue body height

Figure 7 displays the results for our tongue height measures. As in the case of the palatal wide interval, the linear mixed effects regression models for the tongue body height measurements did not benefit from adding random slopes by speaker. Furthermore, the models fit on the basis of the onset of the tongue body movement – the minimum tangential velocity (V_height_onset) or the 20% threshold to peak velocity (V_height_Shawetal) – did not justify adding vowel length and showed no significant differences between the word accents (see Tab. 3). For the measurements taken from landmarks during the lowering (peak velocity) and at the target, *vowel length* as fixed effect was warranted. The model fit on the basis of the peak tangential velocity landmark (V_height_peak) showed a significant difference between the word accents (t = -2.4, p < .05) and an interaction with vowel length (t = 5, p < .001), as well as between the vowel lengths (t = -2.5, p < .05). As for the landmark at tongue body target there was also a significant difference between A1 and A2 (t=-2.6, p < .05), a significant interaction of word accent and vowel length (t = 3.0, p < .05), but no main effect of vowel length. The estimated average tongue body height differences between the word accents were slim: 0.2 mm for the variable V_height_peak, which is beneath the reliable resolution of the EMA. Figure 7 indicates a lot of variation, which we would interpret as a sign of speaker variability. To conclude, the difference in tongue body height between the two word accents did not reach significance at gestural onset. Only during the movement of the tongue and at the target was there a significant difference between the word accents (see Tab. 3).

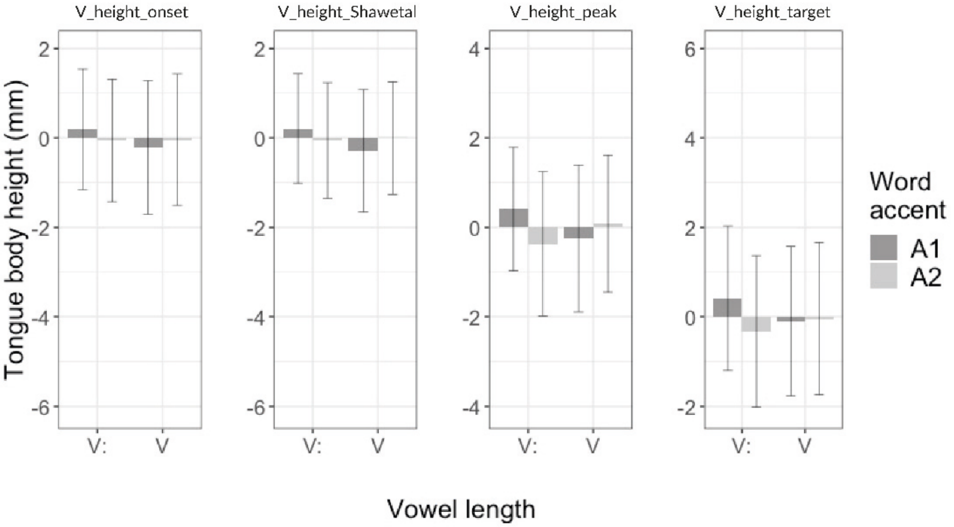


Figure 7. Tongue body height (mean-centered for reasons of clarity) as an effect of word accent and vowel length: at minimum tangential velocity of tongue body onset (V_height_onset), at the 20% threshold of peak vertical velocity (V_height_Shawetal), at peak tangential velocity (V_height_peak), and at minimum tangential velocity of the tongue body target (V_height_target). The ranges of the y-axes have been manipulated as a means of visual support, such that relative tongue position is roughly indicated by the position of 0 on the y-axis.

In order to better understand the significance of the results on the tongue body height given the large variability, a qualitative analysis of the tongue body data was performed. The words were first time-normalized by z-transforming the time points (for details see 2.3), and vertical positions of the tongue body sensor were normalized for each word by z-transforming the positions per word on the basis of all positions in the utterance it occurred in. In Figure 8, regression lines for the vertical tongue body data visualize the differences between the word accents and the vowel lengths. The figure suggests that A2 has a higher vertical tongue body than A1 already during the word-initial consonant, as well as after the tongue body has hit its target (the lowest point) during the vowel.

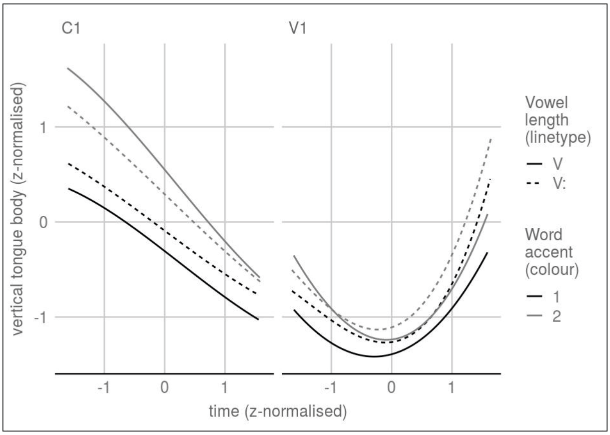


Figure 8. Polynomial regression lines for the normalized vertical tongue body data during the word-initial consonant (C_1) and the vowel (V_1), as an effect of word accent and vowel length.

3.2.4 CV time lag

Finally, we compared models on the time lag measurements between the onset of the bilabial closure (C) and the onset of the tongue body movement (V). Once again, neither of the models justified adding random slopes by speaker. For the two CV time lags that were measured based on lip and vertical tongue body velocity (CV_timelag_Mücketal and CV_timelag_Gao), *vowel length* as a fixed factor was not warranted either. The models showed no significant difference between the two word accents (see Tab. 3). On the other hand, *vowel length* as fixed effect did contribute to the model fit on the CV time lag that was measured based on the acceleration of the lips and the tangential velocity of the tongue (CV_timelag_L&G). The results for this CV time lag measure revealed a significant difference between the word accents (t = -4.8, p < .001), as well as between the vowel lengths (t = 3.7, p < .001), but no interaction between word accent and vowel length.

Furthermore, the estimated average time lag, word accents and vowel lengths combined, was 60 ms based on when the lip and the vertical tongue movements were minimal (CV_timelag_Mücketal), compared to 68 ms for when the movements were in progress (CV_timelag_Gao). In contrast, there was no estimated time lag to speak of (1.4 ms) when measured based on the acceleration of the lips and the tangential velocity of the tongue (CV_timelag_L&G), thus suggesting synchronized CV onsets.

The estimated difference in time lag between the word accents (for short vowels) was 8 ms.

To conclude, although Figure 9 suggests slight differences between the word accents in all three panels, it was only the synchronized CV onsets (CV_timelag_L&G) (bottom panel) that showed a significant difference between the word accents: in A1 the movement towards bilabial closure accelerates before the tongue body movement starts, while in A2 the lips accelerate after the tongue has started to move.

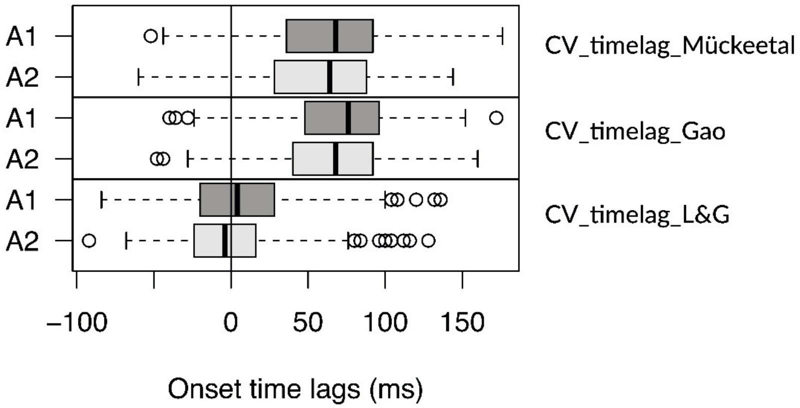


Figure 9. Time lags between bilabial closure onset (0) and tongue body onset (boxes) as an effect of word accent, based on: zero-crossing in the velocity curve for lip aperture and the vertical tongue body movement (CV_timelag_Mücketal), 20% threshold from zero-crossing to peak velocity for lip aperture and the vertical tongue body movement (CV_timelag_Gao), and lip acceleration and minimum tangential velocity of the tongue body (CV_timelag_L&G).

4 Discussion and conclusions

The present study has revealed small but significant spatiotemporal differences in the articulation of a CV syllable onset between the two tonal contexts provided by the Swedish word accents, A1 and A2. The f0 analysis confirmed that the A1 and A2 target words were realized with different tonal patterns as described in the literature, and that the tonal difference is evident and statistically significant already in the word-initial consonant of the CV sequence. In Figure 2, articulatory measurements that showed statistically significant differences between the word accents are marked with an asterisk (*), and they are: the bilabial closure (as measured with maximal acceleration/deceleration landmarks), the palatal wide interval, the tongue body height during tongue lowering and at the target (the tangential velocity maximum and subsequent minimum landmark), and the CV time lag (the lag between onset of lip closure, measured at maximal acceleration, and onset of the tongue body movement, measured at tangential velocity minimum). In the following sections, the results are discussed with respect to each articulator and the coordination between them.

4.1 The lip movements

A few deductions can be derived from the results on the bilabial intervals. First and foremost, the mixed effects regression models revealed a statistically significant difference between A1 and A2 for only one of the six bilabial intervals (the three closure and three release intervals), namely the bilabial closure interval as measured with acceleration landmarks. This result tells us that there is a difference in how the lips are closed for the bilabial constriction in the production of /ma/ in A1 words as compared to A2 words. Measuring the bilabial closure intervals with velocity landmarks did not show this difference. Perhaps this is due to a difference in openness between the lip movements in the two tonal contexts of A1 and A2. From the second order equation we know that acceleration, velocity and position are dependent of each other (see Iskarous, 2017, for an account on a Dynamical Systems Theory for speech). If the timing of velocity-based landmarks for the lip movements does not vary while the timing of acceleration-based landmarks does, it is most likely that the lips are positioned differently in A1 compared to A2. We have already seen a tendency for this in an ongoing study. The work is still in progress but the preliminary results suggest that at around closure onset, A1 tends to display a larger lip aperture than A2. At present, we cannot offer a full explanation for such effects of word accent on lip aperture or lip closing. However, we can speculate that these effects may be linked to the head and jaw movements, which we have already observed to differ between the word accents (Frid et al., 2019; Svensson Lundmark & Frid, 2019). Since the jaw is connected to the tongue, these word-accent effects on, in particular, jaw movements might in turn be explained as a secondary effect of the word-accent specific coordinated movements with the tongue, which is discussed in 4.2 below.

The bilabial release interval differed only when measured with velocity landmarks and only between the vowel lengths, the release occurring faster in the short vowel condition. A possible interpretation of this result is that the production of the initial C is speeded up in preparation for an earlier upcoming post-vocalic coda consonant in the short vowel condition. However, since in Swedish, the quantity contrast also signifies different vowel qualities, the effect on the release could instead also be related to the quality of the vowel, although we cannot offer any explanation for why the lips should release faster in connection with the more front vowel.

To conclude, an effect of the tonal context was only found for the closing interval of the lips, suggesting a difference in lip positioning between A1 and A2, prior to word onset.

4.2 The tongue body movements

Our results suggest effects of the tonal context on tongue body height as well as on the duration of the tongue movement towards the vowel target. The duration of the lowering and retraction of the tongue body movement (the palatal wide interval) differed between the word accents (A2 having the longer intervals) and between vowel lengths (long vowels having longer intervals, unsurprisingly). There was no interaction between these two factors on the duration of the interval, which means that the effect of the tonal accent is independent of the effect by the quantity and quality of the vowel. That is, the vocalic tongue movement is shorter in A1 than in A2 no matter whether the vowel is a short [a] or a long [ɑː]. Furthermore, tongue body height differed significantly between A1 and A2 during the lowering of the tongue (at maximum, or peak tangential velocity) and when the target was reached; for both measures, an interaction with vowel length was found, and even a main effect of vowel length during the lowering of the tongue. However, no effects on tongue body height were found at gestural onset, which was captured by two alternative measures (V_height_onset and V_height_Shawetal). Overall, the results suggest a tendency for a lower tongue body (Fig. 8), and a faster tongue body movement, in A1 compare to A2 throughout the CV sequence.

Effects of lexical tones on tongue body height have been observed in previous studies (Erickson et al., 2004; Hoole & Hu, 2004; Shaw et al., 2016), and it has been argued that such relations can be explained as a result of the physiological connections between the tongue and the larynx (see 1.2 above; henceforth, the tongue-F0 effect). This tongue-F0 effect might also contribute to understanding our results on tongue body height, as we will argue below. However, in order to develop a full explanation of the present findings, we also refer to the notions of the tone gesture (Gao, 2008; see 1.2 above) as well as of *downstream targets* as suggested by Shaw and Chen (2019) within the AP framework.

Our results suggest tongue body height differences between A1 and A2, with a tendency for a lower tongue body in A1 at the vocalic target (i.e. at minimum tangential velocity) as well as at peak velocity on the way towards the target (Tab. 3, Fig. 8). If we start with the effect at the target, then we can conclude that the pattern observed seems to be in line with previous findings on the Mandarin tones (Erickson et al., 2004; Shaw et al., 2016), which have revealed a tendency for a lower tongue in connection with low F0. Figure 1 suggests that the F0 curves for A1 and A2 cross each other at around the temporal midpoint of the acoustic vowel, and Fig. 8 suggests that the vowel target is reached around the same point in time. That is, the expected positive correlation between F0 and tongue body height (i.e. the tongue-F0 effect: lower F0 with lower tongue body) is probably observable from around the time when the vowel target is reached and onwards. However, we do not provide any exact data on the correlations between F0 and tongue body height in this paper. Instead, the conclusion on the connection between F0 and tongue body height drawn here is based on a qualitative comparison of our tongue body results (Tab. 3 and Fig. 8) and our F0 mean curves (Fig. 1). That said, we can, with some caution, conclude that our results for the tongue body height at the vocalic target corroborate the previous findings on the Mandarin tones (Erickson et al., 2004; Shaw et al., 2016), and extend them to lexical pitch accents.

However, before the vowel target is reached (and before the F0 curves cross in Fig.1), the reversed pattern is observed, i.e. a negative correlation between f0 and tongue body: a lower tongue body where f0 is higher, as the falling A1 is still rather high, while the rising A2 is still rather low (see Fig. 1 and Fig. 8). This deviation from the tendency predicted by the tongue-F0 effect implies that our results cannot be fully explained by this effect. However, this deviation is not unexpected. The tongue-F0 effect is only a small effect caused by the interaction of articulators, best observed under controlled conditions, but not a constraint. F0 and tongue height are still largely controlled independently of each other, which is obvious from the fact that different vowels (with different tongue positions) can be produced with high and low F0. Furthermore, negative correlations have been reported in earlier studies, and can be observed, for instance, in the production of emphasis: It has been shown that emphasis tends to be encoded by means of higher F0 and a lower jaw as compared to non-emphatic syllables (Ericksson et al., 2017), that is, by two articulatory strategies that contradict the tongue-F0 effect. The study by Ericksson et al. (2017) suggests that this contradiction is resolved by means of a jaw protrusion that accompanies the jaw lowering with the hypothesized effect of "pull[ing] the hyoid bone forward to counteract the F0-lowering effect of jaw opening" (Ericksson et al., 2017). That is, even if tongue position does not show a positive correlation with F0, other articulators can compensate for it. Hence, deviations from the tongue-F0 effect can not only be explained by means of the independent control of F0, but can also be supported by additional physiological mechanisms, like jaw protrusion.

For the present data, we have not investigated whether such mechanisms might come into play in order to compensate for the higher (i.e. still falling) F0 in A1 and the lower (still rising) F0 in A2 where the tongue body is lower in A1 than in A2 (Fig. 1 and 8). However, similarly as in Ericksson et al. (2017), we can assume that there is some force that triggers a lower tongue body trajectory in A1, irrespective of the F0 trace. This force is, we propose, the need to adjust the tongue body movement in an anticipatory manner to an earlier tongue body target in A1, or a later target in A2. In order to develop this explanation, reference to the notion of *downstream targets* as suggested by Shaw and Chen (2019), as well as to the *tone gesture*, might be useful. The notion of downstream targets suggests that a set of articulatory landmarks beyond the gestural onset (such as the target or the offset) may play a role in gestural coordination. Along these lines, it is possible that, in our data, the vocalic target is coordinated with a specific landmark in a hypothesized tone gesture, capturing a point in time during the fall (A1), or the rise (A2), respectively (see Fig. 1).

If we assume that this landmark is timed later in A2 than in A1, a synchronization with the vocalic target would result in an earlier tongue body target in A1 than in A2, explaining our results for the palatal wide interval (shorter for A1; see Tab. 3). The timing of the tongue body targets for A1 and A2 are illustrated in Figure 10, which displays an example of the vertical tongue body position traces, together with the tangential velocity landmarks. Likewise, Figure 10 shows that the target of the short vowels is, naturally, reached earlier than the target of the long ones.

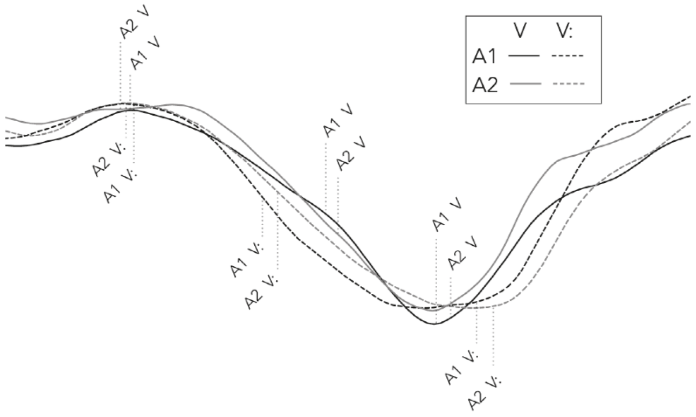


Figure 10. A representative illustration of the timing between the vertical tongue body movement and the landmarks used for the measurement of tongue body height at onset, peak and target. The four lines are aligned at acoustic C onset and represent the y-trace of TB sensors over time in four different target words (A1: *mannen, manen*; A2: *manna, manar*) spoken by one of the speakers. The dotted lines mark the tangential landmarks from which the height measurements are taken, that is, they also include horizontal movement.

At present, we cannot offer an account of what the relevant tonal landmark might be, and whether or not we should refer to it as the target. Note that the tonal structures of A1 and A2 are, in the South Swedish dialect, can be represented as H*L (A1) and L*H (A2), suggesting a low target in A1 and a high target in A2. However, it is not obvious why these targets should be relevant for gestural coordination (and the same would hold for the onset of the tone gesture, often measured in previous studies to calculate onset-to-onset VT time lag, as Gao, 2008, or Mücke et al., 2012). Instead, from a comparison of Figure 1 and Figure 10, we might, tentatively, conclude that the tonal landmark that is coordinated with the tongue body target is to be found during the F0 fall (A1) or rise, respectively. A possible candidate for a relevant F0 landmark (i.e. a downstream target) might be peak F0 velocity (based on the first derivative of the F0 curve), as this landmark would reflect the point in time where the F0 rise/fall is the steepest (cf. Fig. 1), but this hypothesis will need to be tested in future studies.

The account proposed here – a synchronization of vocalic and tonal downstream targets – can now be pursued to explain our results on tongue body height: a higher tongue body in A2 than in A1, both during the tongue lowering (at maximum tangential velocity) and when the vocalic target is reached (see Fig. 8, Tab. 3). In order to reach an earlier or later target, respectively, the tongue will adjust its movement towards its target, which, we assume, might affect the shape of the tongue body trajectory and the timing of our measurement landmark (maximal tangential velocity), and hence also tongue height measured at this landmark. This is, again, illustrated in Figure 10. The figure suggests that the combined vertical and horizontal movement of the tongue starts at approximately the same point, but depending on word accent and vowel length, the tongue movement takes different paths and speeds up differently to reach the target. We can see that the peak velocity landmark is reached later in the curve for the short vowels than for the long vowels. A later point in time during the lowering should correspond to a lower tongue position, which in Figure 10 is evident when comparing the A2 trajectories for the short and the long vowel (but less so for A1, which illustrates the interaction found between word accent and vowel length). Therefore, the difference in tongue body height between long and short vowels measured at the peak velocity landmark (see Tab. 3) is explained by the different timing of this landmark during the lowering. The measured tongue body height differences between the word accents during the lowering of the tongue can then be explained in the same way, as the landmarks in A2 are timed later during the lowering compared to A1.

Finally, the tongue body height differences measured at the vocalic target, which we have related to the tongue-F0 effect above, might likewise be a by-product of the tongue-body movement and how it adapts to the timing of its target: Possibly, since the tongue body lowering is speeded up more and earlier in A1 than in A2 (earlier reached maximum velocity in A1), as suggested by Figure 10, the tongue body finally reaches a slightly lower position in A1 than in A2.

Before concluding this section, we will briefly comment on the relation between the effect of word accent on the palatal wide interval, and previously reported effects on segmental durations. The findings on the palatal wide interval are, generally, in line with a previous articulatory and acoustic pilot study (Svensson Lundmark et al., 2017) and also with earlier acoustic studies of Swedish word accents where segment duration was found to differ between A1 and A2 (e.g., Elert, 1964). However, we cannot easily argue for a connection between the present results on the duration of the tongue movement and previous results on acoustic durations. Durations reported for South Swedish display a difference between A1 and A2 only in the post-vocalic consonant, but not in the vowel itself (Svensson Lundmark et al., 2017). Furthermore, in Svensson Lundmark et al. (2017), we argued, based on a comparison of the duration effects found in South Swedish (Svensson Lundmark et al., 2017) and Central Swedish (Elert, 1964), that the duration difference between A1 and A2 may be an effect of focus, affecting A1 and A2 differently due to the differently timed F0 patterns in A1 and A2. However, in the present study, we have elicited unfocused target words, for which we would not predict any segmental durational effects of word accent. Future studies will have to address the relation between the palatal wide interval and the acoustic duration, as well as the nature of the durational differences between A1 and A2.

To conclude, our results revealed multiple effects (temporal and spatial) of the tonal context (the word accent) on tongue body movements. We have offered a tentative explanation of these effects, in terms of a synchronization between vocalic and tonal downstream targets, and an anticipatory adjustment of the tongue body movement to the timing of the target. Assuming a later tonal target (which is yet to be defined) in A2 than in A1 would directly explain a shorter duration of the tongue lowering in A1 (palatal wide interval). Furthermore, the adjustment of the tongue body movement to the differently timed targets can, we have argued, explain the differences in tongue body height measured during the lowering of the tongue (at maximum velocity) and at the target. A possible contribution of the physiological connection between the tongue and the larynx might also come into play.

4.3 Lip and tongue body coordination

At first glance our results on the CV time lag seem somewhat inconsistent. We found time lags as well as (roughly) synchronized C and V onset, depending on the choice of measure. Furthermore, a small but significant synchronization difference between the word accents was only observed for one measure (CV_timelag_L&G), for which the overall result was a roughly synchronized CV coordination. Table 4 summarizes the results for the CV time lags, obtained with the three alternative measures used in this study, and compares them to results reported in previous studies.

Table 4. CV onset-to-onset time lags obtained with three alternative measures for Swedish (present study), compared to Mandarin (Gao, 2008), and German and Catalan (Mücke et al., 2012). All measurements presented in the table refer to a /ma/-sequence. Data for Swedish are taken from Tab. 3 (intercept estimates from the mixed models based on 19 speakers, summarizing closed and open syllables). The result for Mandarin represents a grand mean across the four tones (pooled across /ma/ and /man/, 7 speakers). Data for German and Catalan are means from different conditions (open vs. closed syllables; broad and contrastive focus; based on 1 speaker per language).

Measure	**Present results: Swedish**		**Previous results**			**Source**
CV_timelag_Gao	**68 ms**		**44 ms**	*(means for individual tones vary between 43-46 ms, n.s.)*	**Mandarin**	Gao (2008)
CV_timelag_Mücketal	**60 ms**		**≈ 5 ms**	(3ms ≤ x ≤7 ms)	**German**	Mücke et al. (2012)
			≈ -2 ms	(-6 ms ≤ x ≤ 3 ms)	**Catalan**	Mücke et al. (2012)
CV_timelag_L&G	**1 ms**	*(difference between A1 and A2 ~8 ms ***)*	*(no data for /ma/ included in Löfqvist & Gracco, 1999)*			

As the presence or absence of time lags between C and V onsets seems to strongly depend on the choice of measure, the findings on the effects of tone gestures on CV coordination in previous studies might in fact be due to methodological decisions (see 1.2), as further discussed below. Even so, our results suggest that there is an effect of word accent on the coordination of C and V (i.e. different time lags for A1 and A2), when measured with the acceleration-based landmark for the lips (CV_timelag_L&G), but that the velocity landmarks do not necessarily capture this effect.

If we were to explain our results in terms of Articulatory Phonology and the c-center effect, this would lean towards the lexical pitch-accent language Swedish behaving more like a tone language than an intonation language, given that we found CV time lags as measured with velocity-based landmarks (as used by Gao, 2008, and Mücke et al., 2012; see Tab. 4). In this sense, our results more closely resemble those obtained for tone languages (Gao, 2008; Karlin & Tilsen, 2015; Hu, 2016). The phonological explanation of the difference in time lags of approximately 20 ms between Mandarin and Swedish (44 ms in Gao, 2008, vs. 68 ms in the present study when measured as in Gao, 2008) could then be that the lexical tones of Mandarin affect the coupling of CV in a different manner than do the tones of the Swedish word accents. However, it would be easier to evaluate these differences if we also had VT time lags available (see 4.4). Furthermore, the actual time lag differences between our study and Gao (2008) could also be explained by factors such as speech rate differences between the studies.

Furthermore, like Gao (2008) and Hu (2016) we did not find any significant differences in CV time lags between tones (A1 and A2 in our case), when using Gao's (2008) measure. This might indicate that the Swedish word accents exhibit identical coupling relations with the consonant and the vowel[2]. On the other hand, using Löfqvist and Gracco's (1999) measure (based on lip acceleration and tongue tangential velocity), we found overall synchronized C- and V-onsets, and at the same time small but significant differences in CV coordination between A1 and A2. This suggests that the present results on the CV time lags are not necessarily explained as an instance of the c-center effect. Instead, the small but significant difference (lips started moving earlier in A1) could, we propose, be a secondary effect caused by the differences in the lowering and the retraction of the tongue body between A1 and A2 (see 4.2). In Figure 10, we have seen an earlier tongue body target in A1, reached through a faster lowering of the tongue than in A2, a lowering which tends to start slightly later in A1 than in A2.

However, we still find an overall CV time lag when measured with velocity-based landmarks (CV_timelag_Gao; CV_timelag_Mücketal). If we at the same time argue

[2] Notably, the model that Gao (2008) derives from her results predicts a difference between tone 4 in Mandarin on the one hand, and tone 1, 2, and 3 on the other, which is, however, not confirmed by her data.

that this time lag is not explained by a c-effect caused by the competition between a consonantal and a tonal gesture onset, then one could argue, that our explanation would also predict CV time lags (when measured with velocity-based landmarks) for non-tonal languages such as Catalan, Italian, and German, although synchronous CV onsets have been reported for these languages by Niemann et al. (2011) and Mücke et al. (2012).

However, we do not believe that a comparison of independent results on CV time lags from different languages is entirely conclusive in the present discussion, for at least three reasons. First, different measures for gestural onsets provide different time lags, but which ones are most valid? An independent evaluation of alternative gesture onset measures, for each articulator, is required. Second, even if studies on different languages use a common measure (as – almost – in the case of Gao, 2008, and Mücke et al., 2012), differences in time lags found for different languages cannot unambiguously be attributed to the presence or absence of lexical tone in these languages, as CV coordination is language specific (see our discussion in 1.2 and, e.g., Fowler & Saltzman, 1993). Third, previous results on CV time lags based on only one (as in Mücke et al., 2012) or a few participants (as in Gao, 2008) cannot even reliably be regarded as representative of the language in question, as a high degree of speaker variability has been attested in previous studies. For instance, Löfqvist and Gracco (1999) reported speaker (and context) dependent patterns of inter-gestural coordination: In their analysis of CV time lags in a VCV sequence, they observed a consistent negative time lag (final V-onset preceding the C-onset) for one their three American English speakers, while the other two displayed both positive and negative time lags depending on vocalic context, with to some extent different effects by vocalic context for the two speakers.

Hence, it is possible that with larger datasets we would indeed find evidence for similar patterns of CV time lags in tonal vs. non-tonal languages, which would involve a high degree of context and speaker dependency. In particular, when the labial landmark is measured with velocity, we might generally observe CV time lags between the lips and the tongue body of approximately 40-60 ms on average, depending on speech rate (for /ma/, as in Gao, 2008, and the present study), as predicted above. Likewise, a general tendency for synchronized CV could be expected when the labial landmark is measured with acceleration (as observed in the present study, and roughly observed in Löfqvist and Gracco, 1999, although their data are not entirely comparable to ours, since they did not include a /ma/-sequence). However, such a comparison would presuppose a normalization for language-specific coordination patterns.

When in time a gesture is made is undoubtedly important for articulatory coordination and it differs with all certainty between different languages. However, the present results suggest the need for caution when using time lags in the present form as a method to measure coordination between C and V in different languages. These measurements can be susceptible to individual differences such as speech rate and, as our study shows, the choice of landmarks has a major impact on the results. A consensus is needed on where an articulatory movement begins and ends, and how it should be measured.

4.4 Limitations and future studies

The discussion so far has revealed various important directions for future research. First and foremost, previous studies on CV coordination, including the present one, should be replicated and validated for tonal and non-tonal languages, using a consistent set of measuring landmarks and a sufficient number of speakers. It is also important to include different vowels, different initial consonants, and different phonotactic and prosodic conditions, as these also have a great influence on CV interactions.

In this study, we have focused on articulatory movements under two prosodic conditions, rather than integrating measurements of tonal events in the articulatory analysis, as done in some previous studies (e.g., Gao, 2008; Mücke et al., 2012). The exclusion of such measures is a limitation, as an analysis of tonal-articulatory temporal coordination would provide further evidence useful for testing the tone gesture hypothesis and the c-center effect (e.g., Gao, 2008). However, we opted to narrow down the scope of this study to a first attempt of establishing effects of the tonal distinction on the CV articulatory movements, that is, tone-induced shifts on CV, not least in order to provide space for a comprehensive data analysis of lip and tongue movements. Furthermore, measuring f0 minima and maxima – in order to enable a comparison with existing analyses by Gao (2008) and Mücke et al. (2012) – would have required a speech material that is not easily designed for Swedish (see 1.4). In addition, as discussed in 4.2, future research needs to establish the tonal (downstream) targets of the hypothetical tone gestures of the South Swedish word accents. Furthermore, the perceptual relevance of f0-extrema and their timing, as traditionally used in segment-to-pitch analyses, including the cited studies on tonal-articulatory alignment, have been questioned (e.g., D'Imperio, 2000; Knight, 2003; Niebuhr, 2007). More valid means of representing f0 in such analysis have been proposed (e.g., Barnes et al., 2008; 2012; Cangemi et al., 2019) which could be used in future studies on tonal-articulatory alignment.

Given the comprehensiveness of the present study, we have excluded jaw movements from the analysis and refer to a parallel study (Svensson Lundmark & Frid, 2019). We have also excluded lip position data but referred to ongoing work in a subsequent study. We should mention that we have not used peak velocity of the lips as a landmark, which is a viable method of measurement used in many articulatory studies on consonants

(Byrd et al., 2005; Cho, 2002; Erickson et al., 2014; Mooshammer et al., 2006). We will follow up and relate to consonantal peak velocity in the work ahead.

Speaker variability was accounted for in our mixed effects models, but a detailed analysis of the phenomenon per se was deemed outside the scope of the present study. However, the number of speakers included in this study would enable a future exploration of individual speaker strategies as a possible source of variability.

The differences for the tongue body and the lips found between the word accents most likely also affect the acoustic space beyond f0. The tongue body movements presumably affect both formants and intensity, while the bilabial release could affect the transition from the consonant to the vowel, which it is concurrent with. This presumed acoustic difference might possibly render the word accent difference perceivable even before the tones are clearly distinguished by virtue of F0. The strong relationship between the tone and the suffix in Swedish word accents suggests that any anticipatory cue in word onsets would be of help to the listener in decoding the morphology, which has already been demonstrated for the tonal pattern realized in the initial (stressed) syllable (e.g., Roll et al., 2010; Roll, 2015; Roll et al., 2017; Söderström et al., 2017). The present study suggests that further anticipatory acoustic cues beyond f0 might come into play, possibly even before the tonal pattern is clearly distinguished. We plan to follow up this study with acoustic analyses, specifically on formant values, and perception tests.

4.5 Conclusions

Previous studies have shown that coarticulatory anticipatory movements are sensitive to phonotactics as well as suprasegmental information, but studies on the coordination with f0 have been sparse. Moreover, there is a certain risk of artificial results for studies involving a small number of speakers. Since our study is based on a larger number of speakers than the previous related studies (nineteen, as compared to between one and seven per language in the previous studies), we are more likely to avoid such artefacts. Furthermore, this is the first comprehensive study investigating the articulatory movements of the lips and the tongue body as a function of the Swedish word accents.

As far as this particular study goes, the two Swedish word accents, A1 and A2, portray different lip movements and different tongue body movements, as well as slightly different coordinated patterns between the lips and the tongue body, presumably as a secondary effect of the tongue body movements. Hence, the coarticulatory patterns of inherent dynamic articulatory movements reflect an effect of f0 context: Our data suggest that different articulators (the tongue body and the lips), reaching for their targets, are affected in a coordinated and anticipatory manner by f0. It is possible that, in languages like Swedish and Mandarin, phonological tones may have the same effect on CV coordination, but because of subtly varied inter-gestural patterns this effect results in systematic differences between phonemic targets, both within and across languages. Consequently, a range of measurement landmarks are needed for any cross-linguistic comparisons, which call for a combination of velocity, acceleration and position analyses of articulatory movements when comparing languages and dialects.

Acknowledgement

The authors gratefully acknowledge Lund University Humanities Lab. Thank you to Prof. Merle Horne for fruitful discussions during the initial steps of the outline of this study. Parts of this work were originally presented by the first author at the Speech Motor Control conference in Groningen 2017, and she would like to express her gratitude for all the comments from conference delegates. Special thanks to Prof. Anders Löfqvist for valuable comments. Thank you, Dr. Joost van de Weijer, for advice on statistics in R. We greatly appreciate the input from three anonymous reviewers, as well as from three additional anonymous reviewers on a previous version of this study. We also thank Prof. David House for an additional, informal review of the pre-final, revised manuscript. And finally, many thanks to all the participants who graciously worked the EMA.

Statement of Ethics

All participants read and signed an informed consent. Consent form and methodology were borrowed from a related research project granted to Dr. Susanne Schötz (fourth author) by the Swedish Research Council (contract no. 2010-1599) and approved by the Regional Ethical Review Board, Lund, Sweden (reference no. 2011/430).

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This study was supported by a Lund University PhD position awarded to the first author. The contribution by Dr. Frid was supported by an infrastructure grant from the Swedish Research Council (Swe-Clarin, 2019–2024; contract no. 2017-00626).

Author Contributions

The study was designed by PhD candidate Malin Svensson Lundmark in collaboration with her supervisors Dr. Susanne Schötz and Dr. Gilbert Ambrazaitis. Data were collected by Svensson Lundmark assisted by Dr. Schötz who also instructed Svensson Lundmark in the data collection procedure using EMA. Data analysis was performed by Svensson Lundmark assisted by Dr. Johan Frid, with a minor contribution by Dr. Ambrazaitis. Data interpretation and discussion was drafted by Svensson Lundmark and further developed in collaboration with Dr. Ambrazaitis. The manuscript was primarily drafted by Svensson Lundmark, in places with assistance by Dr. Ambrazaitis. All authors contributed with critical revisions.

References

Ambrazaitis, G., & Bruce, G. (2006). Perception of South Swedish Word Accents. *Working papers, Lund University, Department of Linguistics and Phonetics, 52*, 5–8. Retrieved from https://journals.lub.lu.se/LWPL/issue/archive

Ambrazaitis, G., Frid, J., & Bruce, G. (2012). Revisiting South and Central Swedish intonation from a comparative and functional perspective. In O. Niebuhr (Ed.), *Understanding prosody – The role of context, function, and communication* (pp. 138–158). DeGruyter.

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2008). Alternatives to f0 turning points in American English intonation. *The Journal of the Acoustical Society of America, 124*, 2497. doi:10.1121/1.4782826

Barnes, J., Veilleux, N., Brugos, A., & Shattuck-Hufnagel, S. (2012). Tonal Center of Gravity: A global approach to tonal implementation in a level-based intonational phonology. *Laboratory Phonology, 3*(2), 337–383. doi:10.1515/lp-2012-0017

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. doi:10.18637/jss.v067.i01

Boersma, P., & Weenink, D. (2018). Praat: doing phonetics by computer [Computer software], Version 6.0.37. Retrieved 3 February 2018 from http://www.praat.org/

Bombien, L., Mooshammer, C., & Hoole, P. (2013). Articulatory coordination in word-initial clusters of German. *Journal of Phonetics, 41*(6), 546–561. doi:10.1016/j.wocn.2013.07.006

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica, 45*(2–4), 140–155. doi:10.1159/000261823

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: an overview. *Phonetica, 49*(3–4), 155–180. doi:10.1159/000261913

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Bruce, G. (2005). Intonational prominence in varieties of Swedish revisited. In S. Jun (Ed.), *The phonology of intonation and phrasing* (pp. 410–429). Oxford: Oxford University Press.

Bruce, G., & Gårding, E. (1978). A prosodic typology for Swedish dialects. In E. Gårding, G. Bruce, & R. Bannert (Eds.), *Nordic Prosody—Papers from a symposium (*pp. 219–228*).* Lund: Gleerup.

Byrd, D. (1995). C-Centers revisited. *Phonetica, 52*, 285–306. doi: 10.1159/00026218

Byrd, D. (1996). Influences on articulatory timing in consonant sequences. *Journal of Phonetics, 24*(2), 209–244. doi:10.1006/jpho.1996.0012

Byrd, D., Lee, S., Riggs, D., & Adams, J. (2005). Interacting effects of syllable and phrase position on consonant articulation. *Journal of the Acoustical Society of America, 118*(6), 3860-3873. doi:10.1121/1.2130950

Cangemi, F., Albert, A., & Grice, M. (2019). Modelling intonation: Beyond segments and tonal targets. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 572–576). Canberra, Australia: Australasian Speech Science and Technology Association Inc. Retrieved from http://intro2psycholing.net/ICPhS/

Chen, W.-R., Whalen, D.H., & Tiede, M.K. (2019). Mandibular contribution to vowel-intrinsic F0. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 152–156). Canberra, Australia: Australasian Speech Science and Technology Association Inc. Retrieved from http://intro2psycholing.net/ICPhS/

Cho, T. (2002). *The effects of prosody on articulation in English.* New York: Routledge.

D'Imperio, M. (2000). *The Role of Perception in Defining Tonal Targets and their Alignment.* Doctoral dissertation, The Ohio State University.

Elert, C. (1964). *Phonologic studies of quantity in Swedish.* Stockholm: Almqvist & Wicksell.

Erickson, D., Honda, K., & Kawahara, S. (2017). Interaction of jaw displacement and F0 peak in syllables produced with contrastive emphasis. *Acoustical Science and Technology, 38*(3), 137-146. doi:10.1250/ast.38.137

Erickson, D., Iwata, R., Endo, M., & Fujino, A. (2004). Effect of tone height on jaw and tongue articulation in Mandarin Chinese. In *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, China 2004* (pp. 53-56). International Speech Communication Association. Retrieved from http://www.isca-speech.org/archive/tal2004

Erickson, D., Kawahara, S., Moore, J., Menezes, C., Suemitsu, A., Kim, J., & Shibuya, Y. (2014). Calculating articulatory syllable duration and phrase boundaries. In S. Fuchs, M. Grice, A. Hermes, L. Lancia, & D. Mücke (Eds.), *Proceedings of the 10th International Seminar on Speech Production (ISSP), Cologne, Germany 2014* (pp. 102-105). Retrieved from http://www.issp2014.uni-koeln.de/wp-content/uploads/2014/Proceedings_ISSP_revised.pdf

Erickson, D., Suemitsu, A., Shibuya, Y., & Tiede, M. (2012). Metrical structure and production of English rhythm. *Phonetica, 69,* 180–190. doi:10.1159/000342417

Fant, G. (1969). Distinctive features and phonetic dimensions. *Quarterly Progress and Status Report, Dept. for Speech, Music and Hearing, KTH Stockholm, 2-3*, 1-18. Retrieved from https://www.speech.kth.se/qpsr/

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech, 36*(2–3), 171–195. doi:10.1177/002383099303600304

Frid, J., Svensson Lundmark, M., Ambrazaitis, G., & House, D. (2019). EMA-based head movements, word accent and vowel length. In *Book of Abstracts MMSYM 2019, the 6th European and 9th Nordic Symposium on Multimodal Communication, Leuven, Belgium 2019* (p. 11). Leuven, Belgium: University of Leuven. Retrieved from http://mmsym.org/wp-content/uploads/2016/09/MMSYM2019-book-of-abstracts-0905.pdf

Gao, M. (2008). *Tonal alignment in Mandarin Chinese: An articulatory phonology account.* Doctoral dissertation. Yale University, New Haven.

Gårding, E., Fujimura, O., Hirose, H., & Simada, Z. (1975). Laryngeal control of Swedish word accents. *Working papers, Lund University, Department of Linguistics and Phonetics, 10*, 53–82. Retrieved from https://journals.lub.lu.se/LWPL/issue/archive

Honda, K., Hirai, H., Masaki, S., & Shimada, Y. (1999). Role of vertical larynx movement and cervical lordosis in F0 control. *Language and Speech, 42*(4), 401–411. doi:10.1177/00238309990420040301

Hoole, P., & Hu, F. (2004). Tone-Vowel Interaction in Standard Chinese. In *Proceedings of International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing, China 2004* (pp. 89–92). International Speech Communication Association. Retrieved from http://www.isca-speech.org/archive/tal2004

Hu, F. (2016). Tones are not abstract autosegmentals. In J. Barnes, A. Brugos, S. Shattuck-Hufnagel, & N. Veilleux (Eds.), *Proceedings of the 8th International Conference on Speech Prosody), Boston, USA 2016* (pp. 302–306). International Speech Communication Association. doi:10.21437/SpeechProsody.2016

Iskarous, K. (2017). The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *Journal of Phonetics, 64*, 8–20. doi:10.1016/j.wocn.2017.05.003

Jong, K. de, Beckman, M. E., & Edwards, J. (1993). The interplay between prosodic structure and coarticulation. *Language and Speech, 36*(2–3), 197–212. doi:10.1177/002383099303600305

Karlin, R., & Tilsen, S. (2015). The articulatory tone-bearing unit: Gestural coordination of lexical tone in Thai. *Proceedings of Meetings on Acoustics, 22*(060006), 1–9. doi: 10.1121/2.0000089

Kawahara, S., Masuda, H., Erickson, D., Moore, J., Suemitsu, A., & Shibuya, Y. (2014). Quantifying the effects of vowel quality and preceding consonants on jaw displacement: Japanese data. *Journal of the Phonetic Society of Japan, 18*(2), 54–62. doi: 10.24467/onseikenkyu.18.2_54

Knight, R. (2003). *Peaks and Plateaux: The production and perception of high intonational targets in English.* Doctoral dissertation, University of Cambridge.

Kroos, C., Hoole, P., Kühnert, B., & Tillmann, H. (1997). Phonetic evidence for the phonological status of the tense-lax distinction in German. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 35*, 17–25. Retrieved from https://www.phonetik.uni-muenchen.de/forschung/publikationen/fipkm/

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. doi:10.18637/jss.v082.i13

Lehiste, I. (1970). *Suprasegmentals.* Cambridge, Massachusetts: M.I.T. Press.

Lehiste, I., & Peterson, G. E. (1961). Some basic considerations in the analysis of intonation. *Journal of the Acoustical Society of America, 33*(4), 419–425. doi:10.1121/1.1908681

Lindblom, B., Agwuele, A., Sussman, H. M., & Cortes, E. E. (2007). The effect of emphatic stress on consonant vowel coarticulation. *Journal of the Acoustical Society of America, 121*(6), 3802–3813. doi:10.1121/1.2730622

Löfqvist, A. (1975). Intrinsic and extrinsic f0 variations in Swedish tonal accents. *Phonetica, 31*, 228–247. doi: https://doi.org/10.1159/000259671

Löfqvist, A., & Gracco, V. L. (1999). Interarticulator programming in VCV sequences: Lip and tongue movements. *Journal of the Acoustical Society of America, 105*(3), 1864–1876. doi: 10.1121/1.426723

Macneilage, P. F., & Declerk, J. L. (1969). On motor control of coarticulation in CVC monosyllables. *Journal of the Acoustical Society of America, 45*(5), 1217–1213. doi:10.1121/1.1911593

Marin, S. (2013). The temporal organization of complex onsets and codas in Romanian: A gestural approach. *Journal of Phonetics, 41*(3–4), 211–227. doi:10.1016/j.wocn.2013.02.001

Mooshammer, C., Bombien, L., & Krivokapic, J. (2013). Prosodic effects on speech gestures: A shape analysis based on functional data analysis. *The Journal of the Acoustical Society of America, 133*, 3565. doi:10.1121/1.4806505

Mooshammer, C., Hoole, P., & Geumann, A. (2006). Interarticulator cohesion within coronal consonant production. *Journal of the Acoustical Society of America, 120*(2), 1028–1039. doi:10.1121/1.2208430

Mücke, D., Nam, H., Hermes, A., & Goldstein, L. M. (2012). Coupling of tone and constriction gestures in pitch accents. In P. Hoole, L. Bombien, M. Pouplier, C. Mooshammer, & B. Kühnert (Eds.), *Consonant clusters and structural complexity* (pp. 205–230). Berlin: Mouton de Gruyter.

Nam, H., Goldstein, L., & Saltzman, E. (2009). Self-organization of syllable structure: A coupled oscillator model. In F. Pellegrino, E. Marsico, I. Chitoran, & C. Coupé (Eds.), *Approaches to phonological complexity* (pp. 297–328). Berlin: Walter de Gruyter.

Niebuhr, O. (2007). The signalling of German rising-falling intonation categories—the interplay of synchronization, shape, and height. *Phonetica, 64*(2–3), 174–193. doi:10.1159/000107915

Niemann, H., Mücke, D., Nam, H., Goldstein, L., & Grice, M. (2011). Tones as gestures: The case of Italian and German. In W.-S. Lee, & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong 2011* (pp. 1486–1489).

Nittrouer, S., Munhall, K., Kelso, J. A. S., Tuller, B., & Harris, K. S. (1988). Patterns of interarticulator phasing and their relation to linguistic structure. *Journal of the Acoustical Society of America, 84*(5), 1653–1661. doi:10.1121/1.397180

Ohala, J. J., & Eukel, B. W. (1987). Explaining the intrinsic pitch of vowels. In R. Channon, & L. Shockey (Eds.), *In honor of Ilse Lehiste* (pp. 207–215). Dordrecht: Foris.

Öhman, S. (1965). On the coordination of articulatory and phonatory activity in the production of Swedish tonal accents. *Quarterly Progress and Status Report, Dept. for Speech, Music and Hearing, KTH Stockholm, 6*, 14–19. Retrieved from https://www.speech.kth.se/qpsr/

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *Journal of the Acoustical Society of America, 39*(1), 151–168. doi:10.1121/1.1909864

Öhman, S., Mårtensson, B., Leanderson, R., & Persson, A. (1967). Cricothyroid and vocalis muscle activity in the production of Swedish tonal accents: A pilot study. *Quarterly Progress and Status Report, Dept. for Speech, Music and Hearing, KTH Stockholm, 8*(2-3), 55–57. Retrieved from https://www.speech.kth.se/qpsr/

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*(2), 175–184. doi:10.1121/1.1906875

Pouplier M. (2012). The gestural approach to syllable structure: Universal, language- and cluster-specific aspects. In S. Fuchs, M. Weirich, D. Pape, & P. Perrier (Eds.), *Speech planning and dynamics* (pp. 63–96). New York, Oxford, Wien: Peter Lang.

R Core Team (2015). R: A language and environment for statistical computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Recasens, D. (2002). An EMA study of VCV coarticulatory direction. *Journal of the Acoustical Society of America, 111*(6), 2828–2841. doi:10.1121/1.1479146

Reinholt Petersen, N. (1978). Intrinsic fundamental frequency of Danish vowels. *Journal of Phonetics, 6,* 177–189.

Roll, M. (2015). A neurolinguistic study of South Swedish word accents: Electrical brain potentials in nouns and verbs. *Nordic Journal of Linguistics, 38*(2), 149–162. doi:10.1017/S0332586515000189

Roll, M., Horne, M., & Lindgren, M. (2010). Word accents and morphology—ERPs of Swedish word processing. *Brain Research, 1330*, 114–123. doi:10.1016/j.brainres.2010.03.020

Roll, M., Söderström, P., Frid, J., Mannfolk, P., & Horne, M. (2017). Forehearing words: Pre-activation of word endings at word onset. *Neuroscience Letters, 658*, 57–61. doi:10.1016/j.neulet.2017.08.030

Shaw, J. A., Chen, W. R., Proctor, M. I., & Derrick, D. (2016). Influences of tone on vowel articulation in Mandarin Chinese. *Journal of Speech, Language, and Hearing Research, 59*(6), 1566–1574. doi:10.1044/2015_jslhr-s-15-0031

Shaw, J. A., & Chen, W.-R. (2019). Spatially-conditioned speech timing: evidence and implications. *Frontiers in Psychology, 10*, 2726. doi:10.3389/fpsyg.2019.02726

Smith, C. (1995). Prosodic patterns in the coordination of vowel and consonant gestures. In B. Connell, & A. Arvaniti (Eds.), *Laboratory Phonology IV: Phonology and Phonetic Evidence* (pp. 205–222). Cambridge: Cambridge University Press.

Söderström, P., Horne, M., & Roll, M. (2017). Stem tones pre-activate suffixes in the brain. *Journal of Psycholinguistic Research, 46*(2), 271–280. doi:10.1007/s10936-016-9434-2

Stevens, K., & Blumstein, S. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas, & J.L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

Svensson, M. (2014). Constant tonal alignment in Swedish word accent II. In N. Campbell, D. Gibbon, & D. Hirst (Eds.), *Proceedings of the 7th International Conference on Speech Prosody, Dublin, Ireland 2014* (pp. 987–991). Retrieved from http://fastnet.netsoc.ie/sp7/sp7book.pdf

Svensson Lundmark, M., Ambrazaitis, G., & Ewald. O. (2017). Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents. In *Proceedings of Interspeech, Stockholm, Sweden 2017* (pp. 3236–3240). International Speech Communication Association. doi: 10.21437/Interspeech.2017

Svensson Lundmark, M., & Frid, J. (2019). Jaw movements in two tonal contexts. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 1843–1847). Canberra, Australia: Australasian Speech Science and Technology Association Inc. Retrieved from http://intro2psycholing.net/ICPhS/

Wieling, M., & Tiede, M. (2017). Quantitative identification of dialect-specific articulatory settings. *Journal of the Acoustical Society of America, 142*(1), 389–394. doi:10.1121/1.4990951

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f0 contours. *Journal of Phonetics, 27*(1), 55–105. doi:10.1006/jpho.1999.0086

Xu, Y. (2013). ProsodyPro — a tool for large-scale systematic prosody analysis. In B. Bigi, & D. Hirst (Eds.), *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013), Aix-en-Provence, France 2013* (pp. 7–10). Aix-en-Provence, France: Laboratoire Parole et Langage. Retrieved from http://www2.lpl-aix.fr/~trasp/Proceedings/TRASP2013_proceedings.pdf

Zhang, M., Geissler, C., & Shaw, J. (2019). Gestural representations of tone in Mandarin: evidence from timing alternations. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 1803–1807). Canberra, Australia: Australasian Speech Science and Technology Association Inc. Retrieved from http://intro2psycholing.net/ICPh

Paper III

JAW MOVEMENTS IN TWO TONAL CONTEXTS

Malin Svensson Lundmark[1], Johan Frid[2]

[1]Centre for Languages and Literature, Lund University, Sweden; [2]Humanities Lab, Lund University, Sweden
malin.svensson_lundmark@ling.lu.se; johan.frid@humlab.lu.se

ABSTRACT

Intra-syllabic articulatory movements have previously been shown to vary depending on tonal context. In this study we concentrate on the jaw and explore its movements in two tonal contexts, i.e. the falling and the rising tone of the Swedish word accents. EMA was used to track the mandible of 19 speakers while they read sentences with target words carrying similar word onsets. Evaluation of the acceleration (second derivative) of the jaw movements revealed three well-defined intervals: jaw opening, jaw open posture and jaw closing. Mixed effects models showed that the jaw opening and the jaw open posture were longer in the rising tone context. Concurrently, spatial data revealed lower jaw trajectories in the falling tone context. Our results resemble that of previous findings on the tongue body. In conclusion, the two tonal contexts induce different jaw movements, presumably because of the involved physical mechanism of two distinct tonal targets.

Keywords: jaw, acceleration, EMA, tone

1. INTRODUCTION

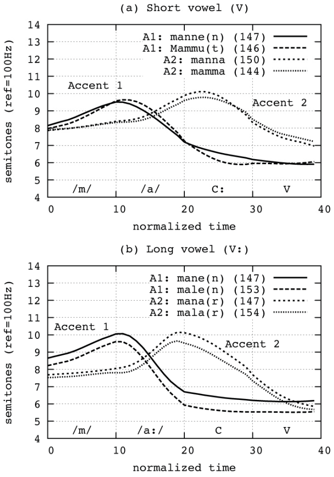
This study explores the movements of the jaw in the two tonal categories of the Swedish word accents (henceforth SWA). SWA are morpho-phonological and sometimes referred to as lexical pitch accents. They are used to differentiate monosyllabic and polysyllabic words, while in disyllabic words they are associated with different suffixes. Phonologically a falling or a rising tone, respectively, is associated with the stressed syllable [1]. This results in an early f0 peak in Accent 1 (A1), and a late f0 peak in Accent 2 (A2) (Fig. 1, from [2]).

What we present here is complementary to previous articulatory findings on the lips and the tongue body (henceforth TB) during the word-initial CV sequence /ma/ [2] in A1 and A2. In [2] a slightly longer lip closure was found in A1 than in A2, while at the same time a longer lowering yet with a higher positioned trajectory of TB in A2 than in A1. In this study we are using the same material as in [2] to perform measurements on the jaw movements.

Previous studies on the coordination of articulatory movements with f0 have mainly focused on Mandarin lexical tones, finding evidence of lip and TB coordination differences between the tones [3], as well as TB height differences [4-6]. In addition, [5-6] found an effect of tone height on the jaw movement: a lower jaw position in Mandarin tone 3 (a low tone). Their results correspond well with the findings by [7] that low tones encompass lower larynx movements and downward movements of the jaw, while higher tones display the jaw moving backward as the tone falls.

These previous findings suggest that the jaw, the tongue, and the lips are highly coordinated with f0 during speech, although the details of the effect of the tonal context are not fully understood. Therefore, in order to get an overall picture an examination of the jaw movements in different tonal contexts is required.

Figure 1: Mean f0 contours of A1 and A2 target words by South Swedish speakers (from [2]): short vowels (a); long vowels (b). Normalized time scale. The vertical lines are segment boundaries.



1.1. Research questions

The goal was to investigate the spatiotemporal role of the jaw during the word-initial CV sequence /ma/ comparing the two tonal categories of the SWA. The research is exploratory and guided by the following questions:

1. Does jaw movement differ between A1 and A2 word-initial CV sequences?

2. Will the jaw follow the same pattern as TB, i.e. have longer intervals in A2?

Because the speech material includes /ma/ sequences with both long and short vowels, and different word-endings, we furthermore suspect the jaw to be affected by these circumstances, e.g. that the closing of the jaw is affected by the post-vocalic consonant, as has previously been found in [8].

2. METHOD

Articulatory data was collected from 19 South Swedish speakers (12 female, $\bar{x}$=40 yrs, sd=12.3 yrs) using a Carstens AG501. Each speaker read leading questions + target sentences from a prompter (presented eight times in random order), an arrangement employed to put a contrastive focus onto the last element in the target sentence. This left the target word in a low-prominence inducing context, hence controlling for possible effects of sentence intonation. In Swedish high prominence is associated with an additional f0 peak or a higher f0 [1, 9].

The original data set consisted of 18 target words, divided into nine word accent pairs with identical stressed syllables (e.g. *bilen-bilar* or *manen-manar*). For this study we only used the four word accent pairs that shared the similar word-initial CV sequence /ma/. However, the target words differed in stressed vowel length and post-vocalic segments (Table 1). The target words were embedded in individual but similarly structured target sentences. The VCV sequence preceding the target words was identical (/ade/). 1191 tokens were recorded.

Table 1: The target words according to SWA and vowel length; long vowels denote open syllables.

	Accent 1	Accent 2
Short vowel	man.nen *the man* mam.mut *mammoth*	man.na *semolina* mam.ma *mom*
Long vowel	mɑː.nen *the mane* mɑː.len *the catfish*	mɑː.nar *urge* mɑː.lar *catfish (pl.)*

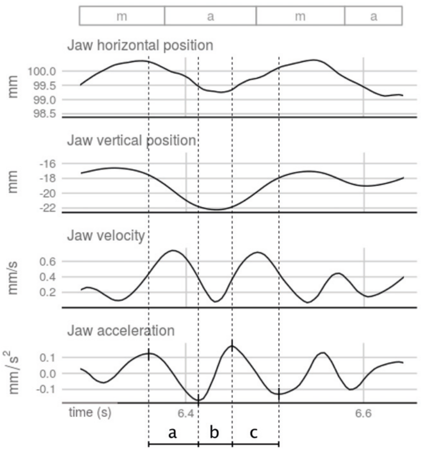
2.1. Procedure

Articulatory data was recorded at 250 Hz, and audio was recorded simultaneously at 48 kHz. For this study we used data from a sensor on the mandible (just below the incisors). In order to correct for head movements, three additional sensors were employed; one behind each ear, and one on the nose ridge. The sagittal angle was not controlled for during the recordings but is not expected to have an effect since we apply the combined x and y dimension. After post-processing including head correction in Carstens software, the articulatory data was transferred to R [11] where it was smoothed using locally weighted regression by the R function *loess* (span=0.1). The first author segmented the acoustic data manually in Praat [12] using ProsodyPro [13]. The TextGrids were further used in R to automatically extract articulatory data from selected time frames of each target word. The time frames were manually adjusted for each speaker using visual cues of the y-trace jaw movements and the tangential velocity profile.

2.2. Measurements

We used the second derivative of the xy trace movement (first derivative of tangential velocity) to locate the moment of maximal acceleration and deacceleration (Fig. 2). Maximal de-/acceleration as onset and offset of an articulatory movement is preferable, since it reflects the damped mass-spring systems of articulatory dynamics ([14] for an account on a Dynamical Systems Theory for speech). In other words, with smoothed acceleration data it is feasible to track speed changes of the articulators, e.g. as they slow down before a complete closure.

Figure 2: Horizontal and vertical jaw trajectories, tangential velocity, and acceleration during an item (*mamma*, A2, segments in boxes). Intervals based on maximal acceleration: jaw opening (a), jaw open posture (b), and jaw closing (c).



After a visual inspection of the jaw acceleration during /ma/, it became evident that there were three well-defined intervals: the jaw opening, the jaw open posture, and the jaw closing (Fig. 2). For this reason, we collected the time points of the maximal de-/acceleration and used them to calculate the three intervals, and also combined them to obtain the total jaw duration. We predicted the temporal measurements to reflect the downward-backward movement, but nonetheless collected data on the vertical position and made a qualitative analysis of the normalized y-trace.

2.3. Analysis

The calculations based on the temporal measurements (the three intervals and the total duration) were statistically tested to see whether they differed between the two tonal categories (A1 and A2). Generalized linear mixed effects models (GLMM) were used to account for sensor placement variation and speaker variability (*speaker* as random effect: random intercept and slope). We added *word* as random effect (random intercept) to avoid certain factors (e.g. word frequency, word endings, or target sentence information structure) enhancing the possibility of a Type I error (=a lower p-value). *Word accent* and *vowel length* were set as fixed effects. We performed a likelihood ratios test for each interval to test whether the additional complexity of vowel length was warranted. The models were run in R using the lme4-package [15]. P-values were obtained by using the lmerTest-package in R [16].

For the qualitative analysis of spatial movements, the words were time-normalized by z-transforming the time points. The jaw sensor vertical positions were then normalized for each word or each speaker by z-transforming positions per word/speaker on the basis of all positions in the utterance it occurred in.

3. RESULTS

3.1 Temporal results

After a model comparison, word accent was fit as the only fixed effect on the jaw opening interval data. The model showed a significant difference between the word accents (*t*=2.41, *p*<.05) (Table 2). Fig. 3A visualizes the small but significant effect of tonal context on the jaw opening. Although there was some variation, A1 had shorter jaw opening intervals than A2 in all four word accent pairs.

The measurements on the jaw open posture interval warranted adding vowel length as a fixed effect to the model. The more complex model revealed a significant effect of word accent (*t*=3.17, *p*<.05), vowel length (*t*=-8.66, *p*<.001) as well as the interaction between them (*t*=-2.76, *p*<.05) on the jaw open posture (Table 2), as illustrated in Fig. 3B. The target words with long vowels (the four bars to the right) have longer and more varied intervals, and in addition seem affected by the word accent as opposed to the short vowels (left bars) (Fig. 3B).

The results on the jaw closing intervals revealed no effect by the fixed effects. Only word accent was warranted, which showed no effect on the jaw closing interval (*t*=0.26, *p*=.8). Error bars in Fig. 3C display some difference between vowel lengths.

Total jaw duration (all three intervals) showed an effect by word accent (*t*=3.17, *p*<.05): the overall jaw movements in A2 are longer (Fig. 3D). The model warranted adding vowel length as well. Table 2 shows the full effects of the model including the interaction.

Fig. 3. Mean intervals in milliseconds (y-axis). A-C: with error bars, all target words divided into SWA pairs with short vowels (left) and long vowels (right). D: all words divided into SWA

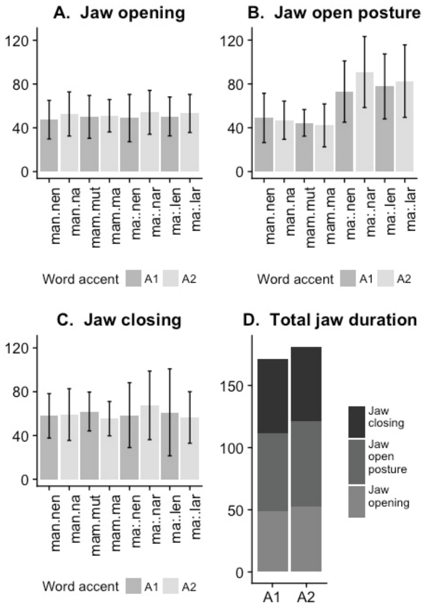


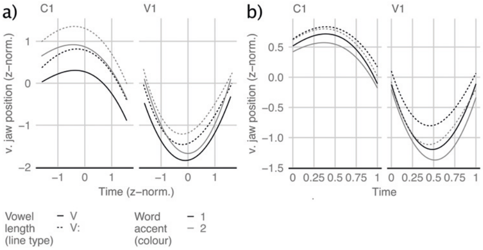
Table 2: GLMM results. Significant effects in bold. Abbr.: intercept (I/C); fixed effects word accent (A1), vowel length (V:), and interaction (A1*V:).

Dep. variable		Est.	SE	df	*t*	p
Jaw opening	I/C	49.37	1.8	19.28	27.38	.000
	A1	3.03	1.26	19.80	2.41	**.026**
Jaw open posture	I/C	75.36	3.47	19.69	21.70	.000
	A1	11.25	3.55	9.17	3.17	**.011**
	V:	-29.19	3.37	8.04	-8.66	**.000**
	A1*V:	-13.07	4.73	7.80	-2.76	**.025**
Jaw closing	I/C	59.21	4.37	23.18	13.56	.000
	A1	0.90	3.46	12.93	0.26	.800
Total jaw duration	I/C	184.39	7.00	23.12	26.47	.000
	A1	20.60	6.49	8.40	3.17	**.012**
	V:	-33.99	6.28	7.43	-5.41	**.001**
	A1*V:	-24.32	8.89	7.45	-2.74	**.027**

3.2 Spatial results

Regression lines on the vertical jaw data were used to visualize the differences between the word accents (Fig. 4). When normalized by word we find a lower jaw trajectory in A1 than in A2. Comparing this to a normalized vertical jaw data by speaker instead reveal a slightly higher jaw in A1. Noticeably, long vowels showed higher jaw trajectories in both word accents.

Figure 4: Regression lines of vertical jaw positions during /ma/ (C1, V1). Normalized by word (a) and speaker (b). Effects: word accent and vowel length.



4. DISCUSSION

The temporal results revealed that the jaw opening and the jaw open posture were shorter in the falling tone (A1) than in the rising tone context (A2). The jaw movements in A1 seemed truncated, as if adjusting for the presence of an early tonal target. On this note, SWA have similar tonal contours, but distinct tonal timing (see Fig. 1). The tonal patterns result in a jaw opening either in a high tonal context (A1) or a low tonal context (A2). Thus, as a shorter jaw opening interval suggests an earlier jaw open posture onset, this might be prompted by the early high tone in A1. Similarly, the open jaw posture was prolonged in A2, as if to cover the rise to the high tone before the jaw closes completely. In other words, the tonal contexts of SWA seem to cooccur with systematic and distinctively timed jaw movements. We suggest that the jaw, and the coordinated articulatory movements (based on findings in [2]), mechanically adjust for the distinct tonal targets of A1 and A2.

Our temporal analysis based on GLMM takes into account the variation between speakers and target words. It also includes dynamical movements, since we use acceleration time points of the downward-backward movements. The results correspond well with earlier findings of shorter TB movement in A1 [2]. In addition, the prolonged jaw movements (total jaw duration) in A2 is consistent with previous acoustic studies on SWA segments [9, 17].

Inspection of the vertical data revealed a lower jaw trajectory for the falling tone context (A1) during /ma/. This could suggest that the shorter intervals in A1 is due to an already low jaw starting position, hence a more open mouth as a possible effect of the high tone context. However, when our data is normalized by speaker the patterns are shifted instead towards a slightly lower jaw for the rising tone (A2). Hence, our two means of normalization either leave out substantial information on the individual target words or speaker variability. Although normalization is fruitful for prominence or lexical frequency effects on segment duration, it leaves out information about the dynamical systems of position, velocity and acceleration, all of which contribute to the shortening or the prolongation of articulatory movements.

The closing of the jaw seems affected by manner and place of articulation for the second consonant, as has partly been reported elsewhere [8], and possibly also by the second vowel, which differs between the target words in our material. However, our data on this is hard to read because of the presence of multiple factors and we leave this for a subsequent study.

The results show that the biggest differences in duration are between phonemically short and long vowels; A1 vs. A2 effects are very small by comparison. However, the open posture intervals of the long vowels appear to differ between A1 and A2, and also display more variance than for the short vowels, which gives some insight into the variability of different vowel lengths. As the longest jaw open posture was found in *manar* (A2) with the highest mean f0 (Fig. 1), prominence (as in high f0) is probably the reason for the longer intervals of this particular target word. However, *manen* (A1), which also carries a high mean f0, has a shorter open posture interval. It seems that prominence has a positive effect on this interval only in late tonal peaks (rising tone), while an earlier peak (falling tone) leads to a negative effect of prominence on the jaw open posture interval. This observation supports our suggestion of a connection between the jaw and f0, though not necessarily reflected in absolute jaw height but rather, as we propose, in timed jaw movements mechanically adjusted to the tonal contexts of the SWA.

5. LIMITATIONS AND FUTURE STUDIES

Analyzing spatial data is intricate. Jaw height has previously been measured at lowest jaw position [5] and at midpoint of the vowel [6], but because of the multiple tonal targets such measurements are not informative for the SWA. Our results suggest that spatial measures need to represent several tonal targets and in addition control for speaker variation.

The data analysis was restricted to the mandible. However, since head posture has been shown to vary depending on tone [6] it should be accounted for in the future. A feasible way is to include data from sensors on the nose ridge or behind the ear.

6. ACKNOWLEDGMENTS

The authors gratefully acknowledge Lund University Humanities Lab. Many thanks to all the participants who graciously worked the EMA.

7. REFERENCES

[1] Bruce, G. 2005. Intonational Prominence in Varieties of Swedish Revisited. In: Jun, S. (ed), *The Phonology of Intonation and Phrasing*, Oxford: Oxford University Press, 410–429.

[2] Svensson Lundmark, M., Frid, J., Ambrazaitis, G., Schötz, S. Consonant-vowel coordination in a lexical pitch-accent language. Submitted.

[3] Gao, M. 2008. *Tonal alignment in Mandarin Chinese: An articulatory phonology account.* (Unpublished doctoral dissertation). Yale University, New Haven.

[4] Shaw, J., Chen, W., Proctor, M., Derrick, D. 2016. Influences of tone on vowel articulation in Mandarin Chinese. *J. Speech Lang. Hear. Res*. 59(6), 1566–1574.

[5] Erickson, D., Iwata, R., Endo, M., Fujino, A. 2004. Effect of tone height on jaw and tongue articulation in Mandarin Chinese. *Proc. Tonal Aspects of Languages* Beijing.

[6] Hoole, P., Hu, F. 2004. Tone-vowel interaction in standard Chinese. *Proc. Tonal Aspects of Languages* Beijing, 89–92.

[7] Honda, K., Hirai, H., Masaki, S., Shimada, Y. 1999. Role of vertical larynx movement and cervical lordosis in f0 control. *Lang. Speech* 42(4), 401–411.

[8] Mooshammer, C., Hoole, P., Geumann, A. 2007. Jaw and order. *Lang. Speech* 50(2), 145–176.

[9] Svensson Lundmark, M., Ambrazaitis, G., Ewald. O. 2017. Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents. *Proc. Interspeech 2017* Stockholm, 3236–3240.

[10] Löfqvist, A., Gracco, V. 1999. Interarticulator programming in VCV sequences: Lip and tongue movements. *J. Acoust. Soc. Am.* 105, 1864 –1876.

[11] R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/.

[12] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer. [Computer program] Version 6.0.37 http://www.praat.org/

[13] Xu, Y. 2013. ProsodyPro — a tool for large-scale systematic prosody analysis. in *Proc. Tools and Resources for the Analysis of Speech Prosody 2013* Aix-en-Provence, 7–10.

[14] Iskarous, K. 2017. The relation between the continuous and the discrete: A note on the first principles of speech dynamics. *J. Phon.* 64, 8–20.

[15] Bates, D., Maechler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *J. Stat. Software* 67(1), 1–48.

[16] Kuznetsova, A., Brockhoff, P., Christensen, R. 2017. lmerTest Package: tests in linear mixed effects models. *J. Stat. Software* 82(13), 1–26.

[17] Elert, C. 1964. *Phonologic Studies of Quantity in Swedish.* Stockholm: Almqvist & Wicksell.

Paper IV

Paper 4: Mutual influence of word-initial and word-medial consonantal articulation

Abstract

This paper examines segment duration in word-initial and word-medial consonants in Swedish di-syllabic words. Results show that the place of articulation of either consonant affects the duration of the other consonant, that is, it has non-local effects on both succeeding and preceding consonants. The findings suggest that the same place of articulation shortens either segment, which may have an articulatory mechanical explanation. The effect of the manner of articulation and gemination were not as convincing, partly due to the design of the study. Furthermore, short consonants and short vowels seemed to vary less than long ones, and word-initial consonants varied less than word-medial consonants.

1 Introduction

This study deals with voiced consonants in CVC sequences that influence each other so that their acoustic duration changes. It is common knowledge that the manner and the place of articulation affect a consonant's segment duration (Cho & Ladefoged, 1999; Harrington, 2013). The explanatory model is naturally derived from articulation. However, small segment duration differences (10-20 ms) have also been detected in consonants though in this case the cause is not due to the inherent consonant, but to non-adjacent segments (Turco & Braun, 2016; Mielke & Nielsen, 2018).

In this study, we investigate some such possible so-called "non-local effects". On the one hand, we investigate whether this non-local effect is present in both the preceding and the succeeding consonants in a CVC sequence. We also divide the effects so that we look at place and manner separately, as well as the effect of gemination.

Furthermore, we include Swedish, a language which, as far as we know, has not been studied from this point of view before. To this end, we have measured the segment duration of both word-initial and word-medial consonants in disyllabic Swedish words.

2 Background

Subtle detailed phonetic differences between consonants may occur as a result of their different articulation. One such example is the variation in voice onset time (VOT) between different stops depending on the location of the articulation (Cho & Ladefoged, 1999). For instance, bilabial stops have been shown to have shorter VOTs than coronal stops (Cho & Ladefoged, 1999; Elert, 1964). To be more precise: alveolar stops are shorter than velar stops (Harrington, 2013). In Swedish, some duration distinction is also found between voiced (which are shorter) and unvoiced consonants (Elert, 1964). Furthermore, manner seems to play a role in Swedish segment durations. Nasals and laterals, for example, are shorter than other manners of articulation (Elert, 1964).

Other segment units are also able to influence a consonant. Perhaps the most well-known examples are those that arise as a result of adjacent segments, or coarticulation. For example, the lip rounding of the vowel is known to affect the nearby preceding consonants, which listeners also use as a perceptual cue (Bell-Berti & Harris, 1979). This effect of lip rounding on nearby segments may be explained by the overlap of consonants and vowels, i.e. the coproduction of gestures (Öhman, 1966; Browman & Goldstein, 1986; Fowler & Saltzman, 1993). Since word-initial consonants overlap the vowel, it is reasonable for the lips to start rounding even before its features arise in the vowel segment.

Another effect which specifically concerns duration may be the relationship between the vowel and the consonant in a VC sequence. For example, vowels have been reported to be shorter before voiceless stops (Cho & Ladefoged, 1999). These and other subtle detailed differences are estimated to be language-specific: thus, for example, in Swedish nasals appear to have a shortening effect on the preceding vowel (Elert, 1964). In terms of place of articulation in Swedish, bilabial consonants seem to lengthen the preceding vowel while in English the opposite case obtains: alveolars lengthen the vowel in comparison with labials (Elert, 1964).

Subtle detailed differences have also been reported to arise as an effect of non-adjacent segments (Turco & Braun, 2016; Mielke & Nielsen, 2018). Such influence cannot as easily be attributed to, for example, the coproduction of gestures or coarticulation. Impact on non-adjacent segments is also sometimes referred to as non-local effects.

2.1 Previous studies on non-local effect

Turco and Braun (2016) studied the non-local effect of consonantal lengthening in Italian. They measured word-initial consonant duration as an effect of whether the following word-medial consonant was a singleton or a geminate. They found that the word-initial segment was longer when followed by a word-medial geminate. Their study partly replicates and corroborates studies of other languages (Hindi and Japanese) in which consonant length contrast is also used (Turco & Braun, 2016). Furthermore, they also compared the geminate words with words with a cluster in which the first of the two consonants were the same as the geminate. They still found a gemination effect, which they ascribed not to a different syllable structure, but to an actual influence of the longer consonant (Turco & Braun, 2016).

Turco and Braun found the same effect regardless of the word-initial consonant; they examined plosives, nasals and fricatives. In a similar study on non-local effects by Mielke and Nielsen (2018) a different approach was used. They did not compare temporal effects such as gemination. Instead they investigated how VOT in voiceless word-initial stops was affected by different types of articulation in succeeding segments. The results revealed that VOT was longer when followed by word-medial liquids in both syllabic and post-syllabic positions. They also found that the voiceless VOT was shorter when followed by post-syllabic voiceless obstruent onsets. This tells us that a) there seems to be word-initial shortening in voiceless stops when it is followed by the same type of articulation (here: aspiration); b) a certain amount of sonorant material may be needed, meaning if you have a voiceless C after the vowel, you shorten the initial C, especially the voiceless part of it.

In Turco and Braun (2016) they speculate on the function of word-initial lengthening, that it might signal the upcoming length contrast, similar to prosodic strengthening. They further imply that the (elsewhere) reported non-local effect by preceding segments (non-local influence in the opposite direction) has a different function: while word-initial lengthening possibly signals upcoming length contrast, word-medial lengthening occurs because of articulatory control (Turco & Braun, 2016). In the present study, our point of departure is that articulatory control and signalling coexist, i.e. they do not contradict each other. Thus, we assume that both word-initial and word-medial lengthening is because of articulatory control and they both also function as signals to the listener.

2.2 Articulatory explanatory models

As shown both by Mielke and Nielsen (2018) and by Turco and Braun (2016), there is a connection between the two consonants on either side of the vowel in a CVC sequence. This relationship may have different causes and explanations depending on how the CVC sequence is composed. In fact, there are various available articulatory explanatory models for the non-local effects depending on the consonantal features. That is, segmental differences that arise due to gemination may not have the same explanation as the ones arising from different manners or places of articulation.

That long consonants have an extended articulation has been shown in previous studies (Smith, 1995; Löfqvist, 2005, 2006, 2007; Türk et al., 2017). The extended constriction of the gesture makes for longer segments (Smith, 1995), which in turn might be due to slower articulatory movements. Several studies indicate that the difference between short and long consonants is related to the velocity of the movements of the articulators (Löfqvist, 2005, 2006, 2007; Türk et al., 2017). However, Swedish speakers have previously not displayed as clear a distinction between long and short consonants as Japanese speakers, which may be due to Swedish complementary distribution (Vː C and VCː, respectively) (Löfqvist, 2005). In any case, the articulatory studies on the geminates suggest that long consonants have a slower motion. This may be part of the explanation as to why word-initial consonants are longer before geminates. Thus, as for the effect of gemination, as in Turco and Braun (2016), a slower articulation would somehow extend across the vowel and into the word-initial consonant. Such an effect, an "anticipatory slowness" if you will, would act as a distinctive feature of a hypothetical bond between the consonants.

There are many possible explanations as to why place or manner of articulation could affect non-local segments. First and foremost, articulation is a cohesion of several articulators. Thus, when the movement of one articulator changes, the relationship with the other articulators changes, and as a result their movement patterns are likely to change. Therefore, there is a high probability that surrounding segments might be affected. Take the jaw as an example. Coronal consonants are generally considered to be associated with large jaw displacement (low jaw) (Mooshammer et al., 2007; Kawahara et al., 2014.). Similarly, there are consonants that give smaller jaw displacements: supralaryngeals, bilabials, and dorsals (Kawahara et al., 2014; Iskarous et al., 2010). Hence, depending on the consonant type, the jaw has different opening degrees. This entails that the production of the nearby vowel is affected by the jaw opening of the adjacent consonant. This may also have an impact on the intrinsic tongue height depending on the vowel type, in its turn, affecting the quality of the vowel.

Mooshammer et al. (2007) witnessed a positive correlation between coronal consonant segment duration and the degree of the jaw opening. The larger the jaw opening, the further the jaw travels spatiotemporally, and the longer the travelled distance, the longer the coronal consonant segment duration (Mooshammer et al., 2007). Because of the angle of the mandible, it is reasonable to assume that constrictions made further back also travel shorter paths, thus making for shorter segments. Nevertheless, the place of articulation is not everything. For example, sibilants usually give rise to a higher jaw due to the constriction, while nasals on the other hand are produced with a lower jaw (Mooshammer et al., 2007; Kawahara et al., 2014). Furthermore, studies on VOT have shown that the speed of the articulator, the sub-glottal and intra-oral pressure, as well as the contact area, affect VOT (Cho & Ladefoged, 1999). Hence, articulation further back in the oral cavity may not entail a shorter segment. Actually, it is only by combining all the factors that we may explain why the bilabial stops seem to be shorter (Elert, 1964; Cho & Ladefoged, 1999), despite being produced at the front of the mouth.

2.3 The purpose of the study

Given that there are many articulatory parameters that may affect the segment duration, it is still unclear how the articulatory properties of a consonant can have a non-local effect across the vowel. Previous acoustic results suggest that there is an articulatory effect on non-adjacent consonants (Turco & Braun, 2016; Mielke & Nielsen, 2018). Since vowels and consonants are believed to be produced separately but are overlapping (Öhman, 1966; Fowler & Saltzman, 1993), a non-local effect would mean that the consonants have a particular bond across the vowel, even though they are not adjacent. If there is such a bond, shouldn't both preceding and succeeding consonants be mutually influenced by changes in articulation? The articulation of a preceding consonant has been shown to affect the following post-vocalic consonant in Estonian (Türk et al., 2017). It is currently unclear whether articulation can be influenced in the opposite direction, that is, by succeeding consonants.

This study contributes to the ongoing research on subtle detailed phonetic variation in segment duration. Our contribution is a) an investigation of a language (Swedish) not previously studied from this point of view; b) results on voiced consonants in which the effects of place, manner and gemination are investigated separately; c) a new approach to this field of work as we explore non-local effects on both word-initial and word-medial consonants. Given previous research, we expect to find an effect in both directions.

Thus, we investigate whether the place and manner of articulation in a consonant has a non-local effect in either direction in a CVC sequence. We also examine whether a word-initial consonant in particular is affected by a word-medial geminate; furthermore, what happens when a word-medial consonant is part of a cluster. Our study can be described as an investigation on the mutual influence of non-local effects.

3 Method

This study follows previous work (Turco & Braun, 2016; Mielke & Nielsen, 2018) as segment duration is used as an indicator of the influence of non-adjacent segments. We measure segment duration of word-initial (C1) and word-medial (C2) consonants. These are either bilabial or dento-alveolar voiced consonants, which are found in disyllabic target words spoken by 18 South Swedish speakers. The material in this study is a subset of the data from an articulatory and acoustic corpus with 21 speakers and approximately 2300 observations (Svensson Lundmark & Frid, 2019; Svensson Lundmark et al., submitted). The subset of the data consists of 18 speakers and of 1127 observations (≈280 observations/set of variables, see section 3.1) and includes only the acoustic data. An analysis of the articulatory data will be presented in a subsequent study.

3.1 Speech material

All the target words are disyllabic Accent 2 Swedish words. Phonetically, this means that there is a tonal rise throughout the vowel with a tonal peak at the syllable border (Öhman, 1967; Gårding & Lindblad, 1973; Bruce, 2005). For those words that have an open stressed syllable (CVː) the tonal peak occurs at the long vowel offset border. For words with a closed stressed syllable (CVC), which in Swedish signifies a short vowel, the tonal peak is located within the coda consonant. In the speech material, target words are placed in target sentences with leading questions, an arrangement employed to put a contrasting focus on the last element in the target sentence. This was to ensure that the target words were spoken in a low prominence inducing context, with no enlarged f_0 excursions or additional tonal peaks that would otherwise normally occur in Swedish phrase accents (Bruce, 1977).

The disyllabic target words are segmentally varied accordingly. All stressed vowels (V1) are open vowels but depending on vowel length they differ in quality: either a short vowel [a] or a long vowel [ɑː]. The second vowel (V2) is always a short unstressed [a]. The segments of interest in this study are the consonants. The subset of the data is limited to voiced consonants. The segments measured in duration (the dependent variables) can either be in word-initial (C1) position - /m n/ - or in word-medial (C2) position - /m l n/. The word-medial consonants may also figure as geminates (/mː nː/). The target words also contain other consonants (see Table 1).

In order to tell whether a non-adjacent articulation has an effect on a consonant we need to compare the same consonant in different contexts. For this purpose, target words that could be minimally paired so that only one consonantal feature differed were matched with each other. This meant that the subset of the data being used (the eight different target words) were further combined into seven different independent variables (Table 1), namely: Place of C1; No C1 (bilabial or no presence of C1); Manner of C1; Place of C2; Manner of C2; Long/short C2 (geminate or singleton); Long/cluster C2 (geminate or cluster)[1]. Hence, the seven independent variables each included two levels as represented by two target words, meaning that some of the eight target words were used more than once (e.g. /ˈmanːa/ was matched in three independent variables). The dependent variables are either word-medial (C2) or word-initial (C1) consonant segment duration (Table 1).

Table 1. Data set and variables

The two-levelled independent variables are the influence of either preceding or succeeding consonants. The dependent variables are the duration of either the word-medial (C2) or word-initial (C1) consonants of the target words.

	INDEPENDENT VARIABLE				TARGET WORDS		DEPENDENT	
			LEVEL 1	LEVEL 2			VARIABLE	
Influence of preceding consonant	1	Place of C1	/m/	/n/	/ˈman:a/	/ˈnan:a/	C2	/n:/
	2	No C1	/Ø/	/m/	/ˈam:a/	/ˈmam:a/	C2	/m:/
	3	Manner of C1	/m/	/b/	/ˈmɑ:lar/	/ˈbɑ:lar/	C2	/l/
Influence of succeeding consonant	4	Place of C2	/n:/	/m:/	/ˈman:a/	/ˈmam:a/	C1	/m/
	5	Manner of C2	/n/	/l/	/ˈmɑ:nar/	/ˈmɑ:lar/	C1	/m/
	6	Long/short C2	/n:/	/n/	/ˈman:a/	/ˈmɑ:nar/	C1	/m/
	7	Long/cluster C2	/n:/	/mn/	/ˈnan:a/	/ˈnamnar/	C1	/n/

3.2 Procedure

Recordings of the corpus took place in the Humanities lab at Lund University. Kinematic and acoustic data were recorded simultaneously. The 21 South Swedish speakers (12 female, x=40 yrs, sd=12.3 yrs) were recorded with sensors glued on their articulators and while attached with cords to an ElectroMagnetic Articulograph, the Carstens AG501. Each speaker read the leading questions + target sentences from a prompter (presented eight times in random order). Articulatory data was recorded at 250 Hz and audio was recorded simultaneously at 48 kHz using an external condenser microphone (t.bone EM 9600). For this study we only used the sound recordings.

The author segmented the acoustic data manually in Praat (Boersma & Weenink, 2018) using ProsodyPro (Xu, 2013). Segment boundaries were established by examining formant transitions of F_1 and F_2. When in doubt, segmentation was led by perceptual judgement.[2] The textgrid output was further analysed in R (R Core Team, 2015).

[1] The whole cluster /mn/ has been annotated as C2 even though it consists of two consonants.

[2] The formant transitions between vowels and nasals were clear and easily segmented. Word-medial /n/ and /m/ were slightly more difficult to judge than the word-initials. The segment boundary between the vowel and /l/ was most difficult and sometimes proved to require listener assessment.

3.3 Analysis

First, the distribution of segment durations was analysed. Since the histograms showed that most of the data were skewed to the right, the raw data were transformed into log scales. These were performed in R with log base 2. Log normal distribution has previously been shown to be useful for segment duration analysis (Rosen, 2005). Raw data over the C1, V1 and C2 segments are also analysed separately.

The log data on C1 and C2 duration, were statistically tested to see whether they differed as a result of the two-levelled independent variables. Generalized linear mixed effects models (GLMM) were used to account for speaker variability (speaker as random effect: random intercept and random slope). To represent the two levels of each independent variable, "word" was set as fixed effects. Hence, the models were established as: [C duration] ~ word + (word | speaker). In addition, we performed a likelihood ratio test to test whether the additional complexities on speakers (random slope) were warranted (following Wieling & Tiede, 2017). The models that were ultimately used can be seen in Tables 2. The models were run in R using the lme4-package (Bates et al., 2015). P-values were obtained by using the lmerTest-package in R (Kuznetsova et al., 2017).

4 Results

The results are presented in three sections. The first section (4.1) contains the results on the influence on the word-medial (C2) duration. Section 4.2 contains the results on the influence on the word-initial (C1) duration. Each section concludes with a summary of the results. Following an interim discussion, the final section (4.3) presents raw duration data on C1, V1 and C2 of all eight target words.

4.1 Influence of preceding segment - word-medial duration

The results on the place of articulation (*variable 1*) suggest that the duration of the word-medial segment is affected by the preceding C1. The GLMM model on the log data shows that C2 is shorter in /ˈnanːa/ (level 2) than in /ˈmanːa/ (level 1) (t=-5.8, p<.001) (Table 2).The distribution spread is about 100 ms, but there seem to be some high outliers (Fig. 1). Furthermore, the distribution is right skewed with a mean value just under 100 ms for /ˈnanːa/, while about 110 ms for C2 in /ˈmanːa/ (Fig. 1).

In *variable 2*, C2 duration seems to be affected by whether there is an occurrence of a word-initial consonant (/m/) or not (t=-6.7, p<.001) (Table 2). When there is no C1 (as in /ˈamːa/, level 1) we find a word-medial consonant lengthening of about 30 ms

(Fig. 2). The spread of data is about 80 ms for /ˈmamːa/ and 150 ms for /ˈamːa/, which also displays a bi-modal distribution (Fig. 2).

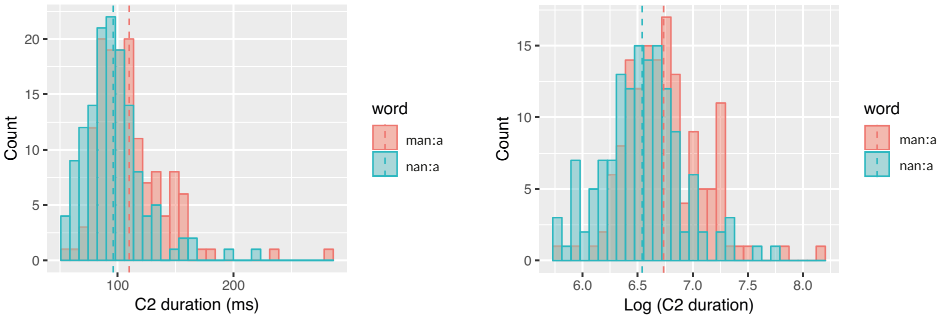


Figure 1. Place of C1 (*variable 1*).
Distribution of C2 segment duration (left) and log data (right). Mean values marked with a dashed line.

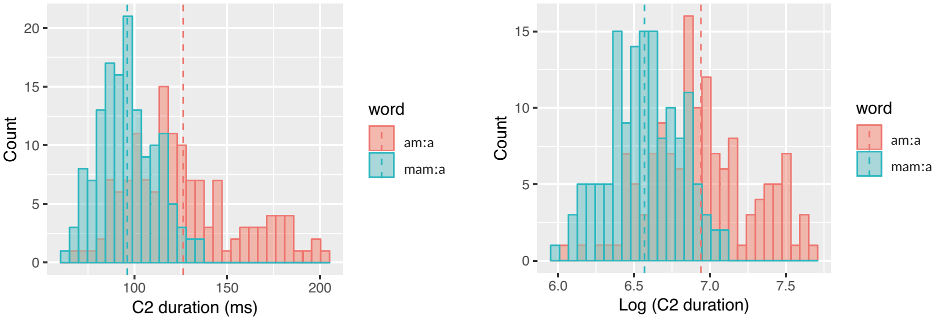


Figure 2. No C1 (*variable 2*).
Distribution of C2 segment duration (left) and log data (right). Mean values marked with a dashed line.

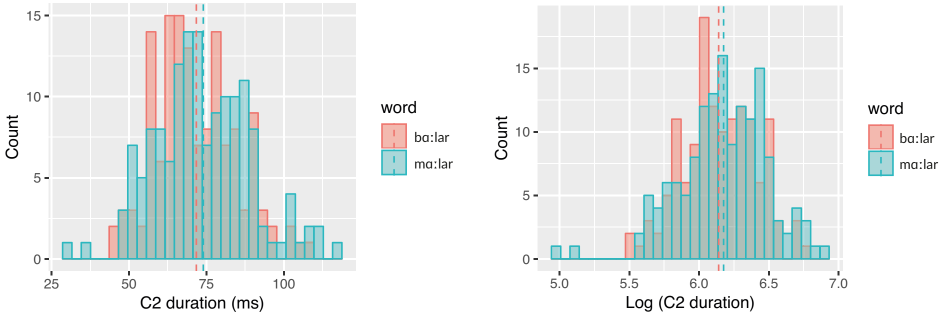


Figure 3. Manner of C1 (variable 3).
Distribution of C2 segment duration (left) and log data (right). Mean values marked with a dashed line.

The manner of the word-initial consonant does not seem to have an effect on the word-medial C2. The peaks are centred and aligned, with a spread of about 80-100 ms (Fig. 3). The difference does not reach significance (t=1.49, p=.137) (Table 2).

Table 2. Results on log data (consonant segment duration).
Results from the GLMMs, with estimates on the effect by word. A less complex model fitted variables 1, 3 and 4.

DEP. VARIABLE		MODEL		ESTIMATE	SE	DF	T-VALUE	P-VALUE
1	C2 /n:/	~ word + (1 | speaker)	Intercept	6.732	0.056	21.560	119.970	.000
			/n/	-0.193	0.034	264.310	-5.772	<.001
2	C2 /m:/	~ word + (word | speaker)	Intercept	6.940	0.074	17.969	94.375	.000
			/m/	-0.373	0.056	18.019	-6.674	<.001
3	C2 /l/	~ word + (1 | speaker)	Intercept	6.140	0.044	22.750	138.66	.000
			/m/	0.042	0.028	268.470	1.49	.137
4	C1 /m/	~ word + (1 | speaker)	Intercept	6.487	0.062	19.050	103.955	.000
			/n:/	0.100	0.021	266.040	4.797	<.001
5	C1 /m/	~ word + (word | speaker)	Intercept	6.572	0.051	17.918	128.361	.000
			/n/	0.064	0.030	18.055	2.189	<.05
6	C1 /m/	~ word + (word | speaker)	Intercept	6.637	0.067	18.063	99.037	.000
			/n:/	-0.050	0.060	18.152	-0.826	.42
7	C1 /n/	~ word + (word | speaker)	Intercept	6.663	0.051	18.307	131.870	.000
			/n:/	-0.053	0.049	18.075	-1.084	.293

4.1.1 Summary of results on word-medial duration

The results on the influence of the word-initial consonant seem to show that when C1 and C2 hold the same place of articulation, C2 is shortened. In *variable 1*, this is shown by the fact that long /n:/ is shorter when preceded by /n/ than when preceded by a bilabial /m/ (Fig. 4).

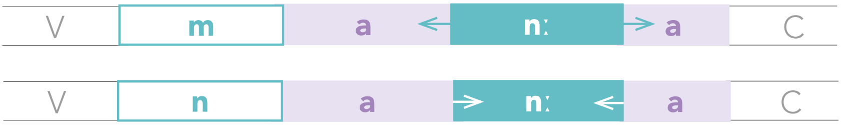


Figure 4. Influence of of preceding consonant: Place of C1 (*variable 1*).
When C1 and C2 has the same place of articulation the duration of C2 /n:/ is shorter.

The influence of same place of articulation might also be the case for the comparison in *variable 2* (No C1). The dependent variable /m:/ is longer in /ˈam:a/ than in /ˈmam:a/ (Fig. 5). Most of our speakers display word-initial glottalization because of the V-V context, in many cases even a full glottal stop. In other words, in words like /ˈam:a/ there is a glottal gesture as a placeholder for C1, in which case it is too a different place of articulation compared to the bilabial. Hence, just like in *variable 1*, the existence of a C1 with the same place of articulation shortens the C2 segment.

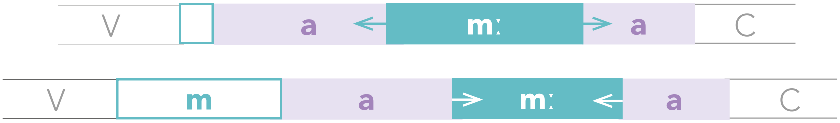


Figure 5. Influence of preceding consonant. No C1 (*variable 2*).
When C1 and C2 has the same place of articulation the duration of C2 /m:/ is shorter.

The manner of articulation (*variable 3*) influences the word-medial segment in our data (Fig. 6). Noticeably, neither /b/ nor /m/ has the same manner of articulation as C2.

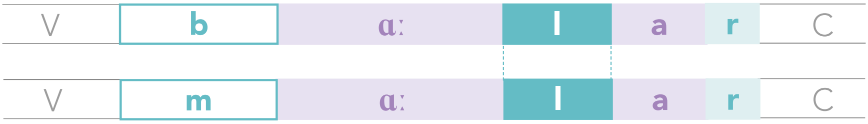


Figure 6. Influence of preceding consonant. Manner of C1 (*variable 3).*
When C1 is nasal the duration of C2 /l/ is slightly shorter, but the difference does not reach significance.

4.2 Influence of succeeding segment - word-initial duration

In *variable 4*, the place of articulation of the word-medial consonant seems to affect the duration of C1 (t=4.8, p<.001) (Table 2). Figure 7 shows a right skewed distribution, where the mean duration in /ˈmamːa/ (level 1) is about 10 ms shorter than in /ˈmanːa/ (level 2). The spread of the data is about 100 ms (Fig. 7).

The manner of word-medial consonant (*variable 5*) also influences the word-initial /m/ (t=2.2, p<.05) (Table 2). The peaks are centred and aligned, with a spread of about 100 ms, although slightly more spread for /ˈmɑːnar/) (Fig. 8). However, with a mean of approximately 100 ms, a word-initial /m/ seem to be 5-10 ms shorter when followed by an /l/ than when followed by an /n/ (Fig. 8).

In *variable 6* there is no difference in C1 duration depending on whether C2 is a geminate or a singleton (t=-0.8, p=.42) (Table 2). The word-initial /m/ in /ˈmɑːnar/ (level 1) even seems to be longer than in /ˈmanːa/ (level 2), according to Figure 9. Distribution is right skewed for /ˈmanːa/ while rather centered for /ˈmɑːnar/, both with a mean of about 100 ms and a spread of data of 120 ms (Fig. 9).

We do not find any effect on the word-initial consonant (*variable 7*) depending on whether C2 is a cluster or not (t=-1.1, p=.293) (Table 2). The distribution is right skewed and the spread of data about 120 ms (fig. 10).

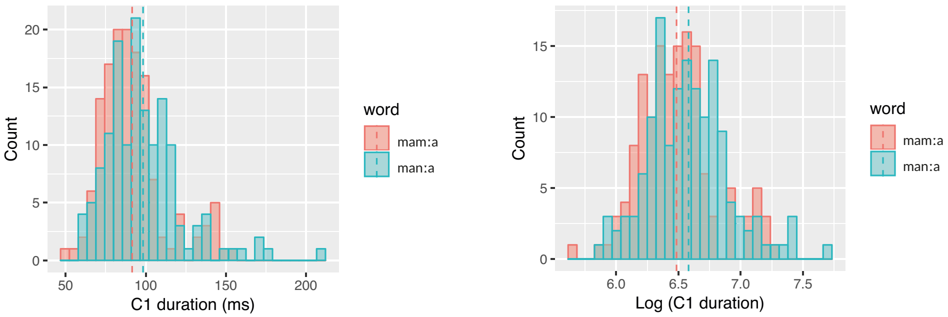


Figure 7. Place of C2 (*variable 4*).
Distribution of C1 segment duration (left) and log data (right). Mean values marked with a dashed line.

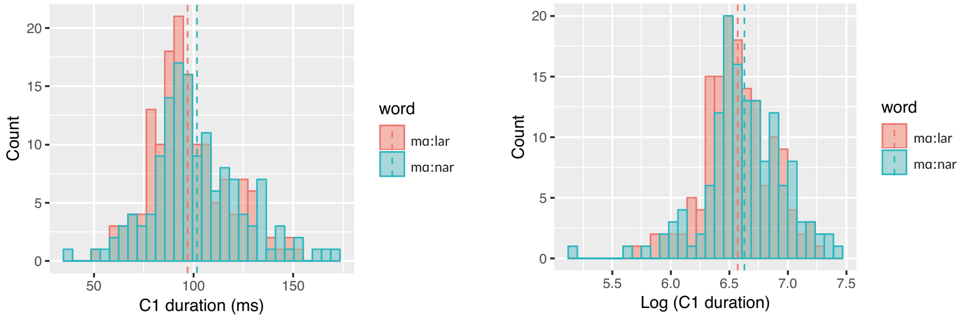


Figure 8. Manner of C2 (*variable 5*).
Distribution of C1 segment duration (left) and log data (right). Mean values marked with a dashed line.

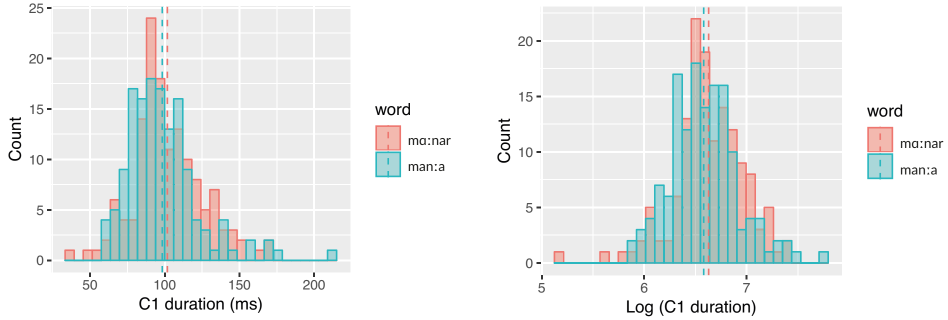


Figure 9. Long/short C2 (*variable 6*).
Distribution of C1 segment duration (left) and log data (right). Mean values marked with a dashed line.

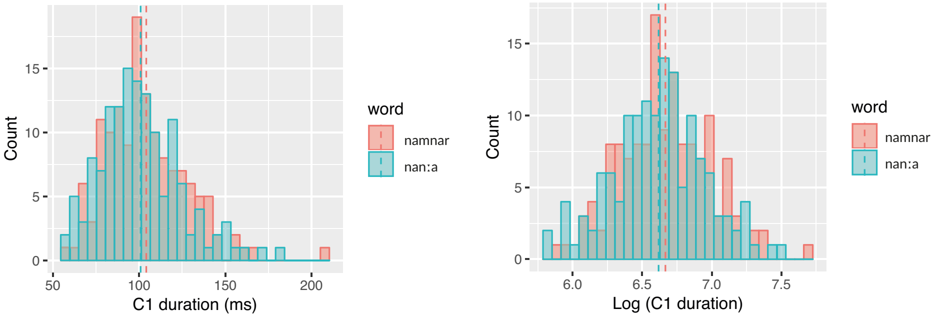


Figure 10. Long/cluster C2 (*variable 7*)
Distribution of C1 segment duration (left) and log data (right). Mean values marked with a dashed line.

4.2.1 Summary of the results on word-initial duration

Again, we find a shortened segment when the place of articulation in C1 and C2 is the same, in this case /m/ (Fig. 11). This suggests that no matter the direction of influence, C1 and C2 will be shorter if the place of articulation is the same. That is, the place of articulation has a non-local effect in both directions.

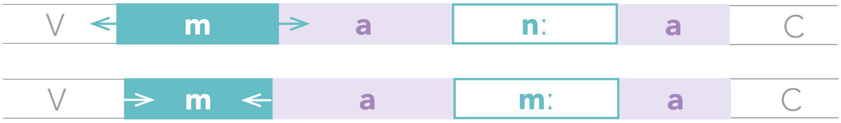


Figure 11. Influence of succeeding consonant. Place of C2 (*variable 4*).
When C1 and C2 has the same place of articulation the duration of C1 /m/ is shorter.

The manner of articulation influences the preceding segment (Fig. 12). Hence, in our example same manner of articulation – here: nasal – seems to lengthen rather than shorten the segment.

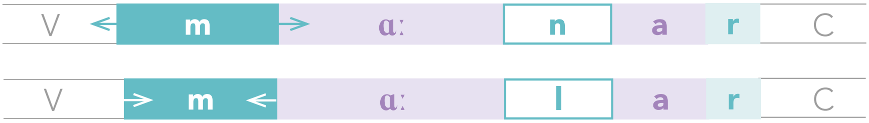


Figure 12. Influence of succeeding consonant. Manner of C2 (*variable 5*).
When C1 is lateral the duration of C1 /m/ is slightly shorter.

We find no gemination effect on the word-initial consonant (Fig. 13). In fact, the results suggest that the word-initial segment in /ˈmɑːnar/ might even be longer than when there is a geminate following. It is likely that other effects have played a role and affected the results.

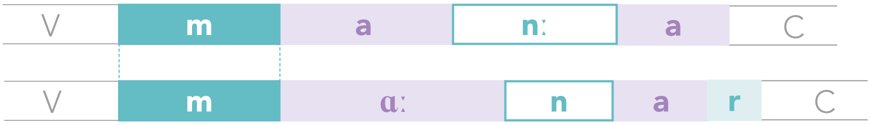


Figure 13. Influence of succeeding consonant. Long/short C2 (*variable 6*).
The effect of gemination of C2 on the duration of C1 /m/. No lengthening/gemination effect found.

Because we already found that the place of articulation has an effect in both directions, we would expect the same pattern in *variable 7* comparing a long consonant to a cluster. That is, /ˈnanːa/ should contain a shorter word-initial /n/ since C1 and C2 have the same place of articulation, while in /ˈnamnar/, C1 and the first C2 in the cluster (/mn/) does not have the same place of articulation. However, we found instead that the word-initial consonants did not differ (Fig. 14).

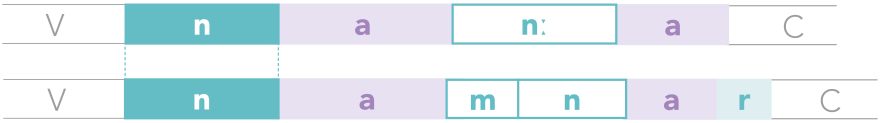


Figure 14. Influence of succeeding consonant. Long/cluster C2 (*variable 7*).
No lengthening effect found on the duration of C1 /n/.

4.2.2 Interim discussion

The distribution of the data in the different variables is similar; right skewed with a spread of 80-120 ms. However, some words deviate from this pattern. For example, we see a slightly more centred distribution in the words included in *variables 3, 5* and *6*, which investigate the effect of the manner of articulation, and the effect of gemination. Common to these words is that they consist of an open stressed syllable followed by a short post-syllabic consonant., i.e. a CV:C sequence. This is in comparison with the words that show a right skewed distribution, which are CVCː sequences, i.e. words with a stressed short vowel and a long consonant.

It is generally known that for Swedish words with a stressed long vowel (V:), the vowel and the subsequent consonant are extended at a higher prominence level (Elert, 1964; Heldner, 2001). Similarly, for words with short vowels in stressed syllables (CVC) this extension is non-linear, i.e. the post-vocalic consonant is more extended than what the short vowel (V) is at a higher prominence (Heldner, 2001). Thus, an almost inverse VC ratio.

The reason for the different prominence ratios between the V and C segments may be due to the tonal patterns that arise when the Swedish word accents become more prominent (see for example Elert, 1964; Gårding & Lindblad, 1973; Bruce, 1977). Thus, the fact that the distribution of our data seems to differ between CVːC and the CVCː words may be due to differences in prominence. To evaluate that possibility, and to further explore the VC ratio, we take a closer look at the raw data of segment duration of the CVC sequences.

4.3 Raw duration data

Figure 15 shows the duration of C1, V1 and C2 in the eight target words used in this study (reminder: only eight target words since some of the words are used multiple times in the seven variables). As evident in Figure 15, the target words with long vowels naturally have longer V1 duration. However, there are small differences between the long vowel words; /ˈbɑːlar/ and /ˈmɑːnar/, have slightly longer V1 than /ˈmɑːlar/ (Fig. 15). As /ˈbɑːlar/ appear to have shorter C2 segments, the longer V1 duration in this particular word might be due to a deviating VC ratio rather than an effect by prominence.

However, the target word /ˈmɑːnar/ stands out because it has both longer V1 and C2 than the other CV:C words (Fig. 15). Thus, in this case it may actually be a possible effect of prominence which could conclude in a Type 1 error in *variable 5* (the effect of manner of C2). Likewise, this presumed prominence effect would also potentially affect the results of *variable 6*, where a gemination effect was not found as /ˈmanːa/ did not have the expected longer C1 than /ˈmɑːnar/.

Notably, the word /ˈmamːa/ has a slightly shorter V1 duration (Fig. 15). This is presumably due to higher informativity. In this subset of the data, the word /ˈmamːa/ ("mother") is the only one with a high lexical frequency.

Of the CVCː sequences where C2 is measured, the target word /ˈamːa/ stands out the most with a C2 duration mean value of 125 ms (apart from the cluster /mn/ which is really two consonants) (Fig. 15). This could be an example of a non-linear prominence extension (Heldner, 2001). However, the longer C2 in /ˈamːa/ may not only be due to prominence, but also an effect of the absence of C1. Since, apparently, /ˈamːa/ displays a C1 duration after all, as seen in Figure 15, consisting of a full glottal stop or a glottalization as a placeholder for C1 (which the analysis work could attest to). Thus, a possible reason for the bi-modal distribution in /ˈamːa/ (Fig. 2) might be the different strategies by the speakers to handle a word-initial V-V context (as a result of the previous word): a full glottal stop or glottalization.

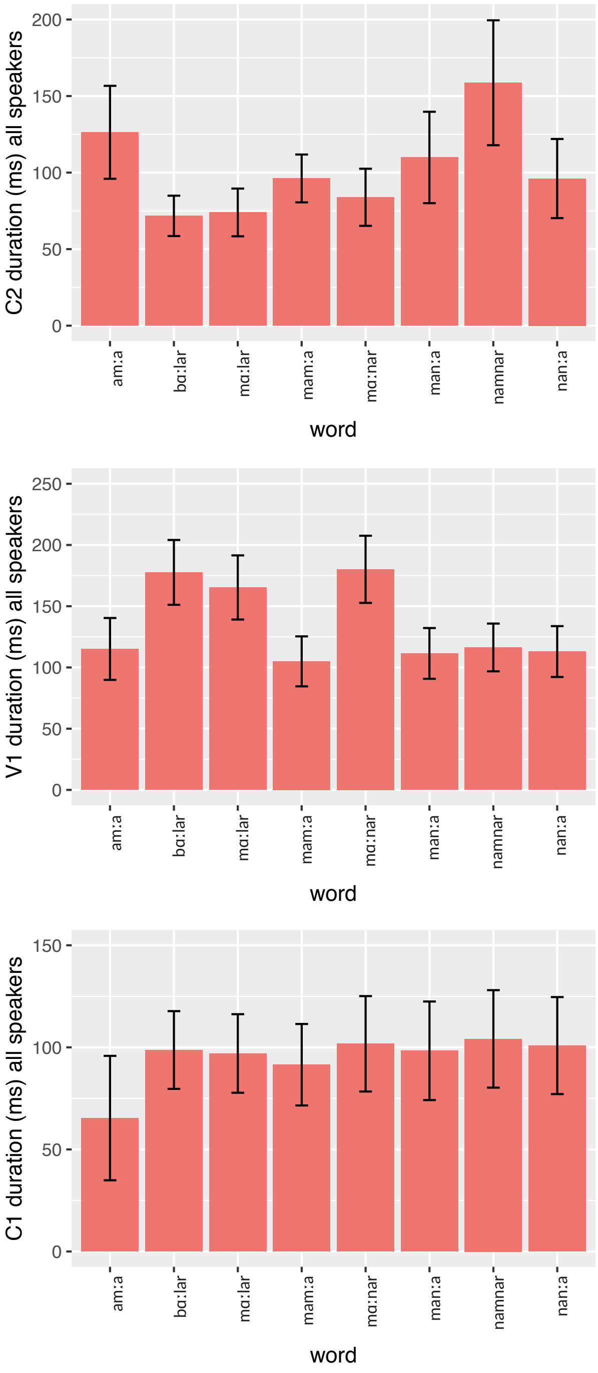


Figure 15. Segment duration of C1, V1 and C2 (from the bottom up).
The word-initial segments of the eight target words used in the study. The bar graphs show mean values and the error bars represent standard deviation.

Furthermore, /ˈamːa/ has a similar V1 duration as the others CVCː words, which indicates that it might not be more prominent after all. In fact, the short vowels have about the same duration (105-115 ms) regardless of consonant combinations both before and after the vowel (Fig. 15). Hence, the ratio of C1 and C2 seems to have some significance, since if you add up the duration of C1 and C2, you would get similar results in e.g. /ˈamːa/ and /ˈmamːa/. However, this only applies to short vowels: according to Figure 15, the long vowels vary more than the short ones. The same can be said about the consonants: long C2s vary more than both short C1s and short C2s (Fig. 15).

5 Discussion

We will first discuss the results of the place of articulation (*variable 1, 2* and *4*). We then move on to the rather inconclusive results on the influence of manner and geminate.

5.1 Mutual influence of the place of articulation

The main finding in our study is that the place of articulation has an effect on both preceding and succeeding consonants. The comparison of the place of articulation has included bilabial and dento-alvolar nasal stops. Hence, we cannot corroborate the results on aspiration by Mielke and Nielsen (2018) who were looking at VOT in voiceless stops. However, in their study they suggested that because there was aspiration in both consonantal positions, the same place of articulation had an influence on the word-initial stop (Mielke & Nielsen, 2018). Our study shows that the same place of articulation also has an effect for nasal stops, that is, the influence by the same place of articulation is not limited to aspiration. In addition, we found not only that the place of articulation influenced the preceding word-initial consonant as in Mielke and Nielsen (2018), but that it also had an effect on the succeeding word-medial consonant, that is, in both directions, what we refer to as a mutual influence of the place of articulation.

In detail, what we found was that the place of articulation shortens both consonant segments. It does not seem to matter where the articulation takes place. What is significant is that it is the same place of articulation. Take a word like /ˈmanːa/ (Fig. 16, second row) where C1 and C2 differ in respect of the place of articulation: C1 is bilabial and C2 is dento-alveolar. If we change either C1 or C2 so that both consonants have the same place of articulation, the phoneme segment duration of the other consonant is affected. The place of articulation – bilabial or dentoalveolar - does not seem to be important: when both consonants in a CVCV sequence have the same place of articulation the consonantal segments are shortened.

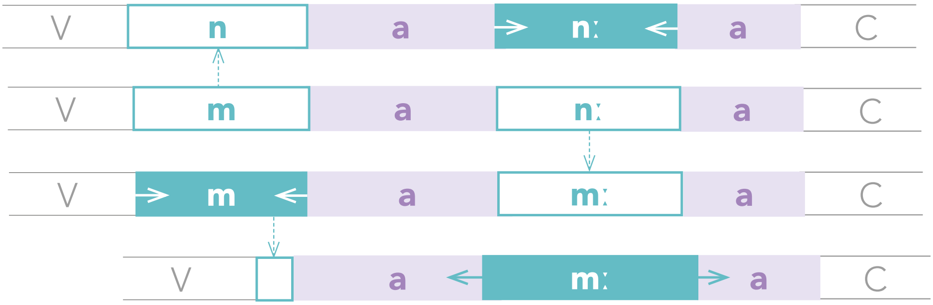


Figure 16. Mutual influence of the place of articulation.
If either consonant changes its place of articulation in the word /ˈman:a/ (the second row), the other consonant is shortened. The same applies when the place of articulation is changed to a glottal constriction (No C1, last row).

This might also be the case for other places of articulation than bilabial and dento-alveolar. The last row in Figure 16 describes what happens when we change the place of articulation of C1 in /ˈmam:a/ to presumed glottalization in /ˈam:a/. Even though there is no C1 in /ˈam:a/ most speakers still produced word-initial glottalization, and sometimes what we presume is a complete glottal stop. That is, they still produced a constriction. More importantly, this constriction had a different place of articulation than the bilabial C2, which is why, presumably, we again find a longer C2 in /ˈam:a/ than in /ˈmam:a/.

Furthermore, this connection of C1 and C2 might be further confirmed by the fact that C2 duration of /ˈam:a/ is both longer and more varied than the other C2s (Fig. 15). Thus, the bimodal distribution of C2 in /ˈam:a/ (Fig. 2) might be a result of the different ways in which the speakers handle the V-V context in C1, i.e. how glottalization is produced. Thus, not only is there a contact between C1 and C2 across the vowel: it also appears to be proportional, controlled by a C1-C2 ratio. A C1-C2 ratio indicates that the short vowel duration is important to maintain. The prolonged C2 in the word with no C1 (/ˈam:a/) therefore reflects the necessity of keeping the short vowel short. The same thing, though the other way around, may apply to long vowels. In Figure 15 it can be seen that the long vowels vary more in length, possibly due to the prominence as already mentioned, but also possibly due to the fact that they are not limited by their length. Thus, long vowels can in principle be any length, while short vowels must be short.

5.1.1 An articulatory explanation model for mutual influence of place

What happens in practice when the place of articulation is the same in both consonantal positions? The effect might be dependent on where the articulation is. In this section we will discuss different articulatory explanations, mainly based on the target words found in *variables 1* and *4*; /ˈmamːa/, /ˈmanːa/ and /ˈnanːa/.

Why would the place of articulation have an effect? The answer lies in how the articulation differs. First and foremost, there is an obvious difference between /m/ and /n/ which is, of course, that their "main" articulator is the lips and the tongue tip, respectively. In terms of inter-articulatory cohesion, /n/ and /m/ presumably both have a time-locked connection between their articulator and velum. Hence, the inter-articulatory cohesion between the two should be somewhat similar. Moreover, both articulators have a small mass, and one can assume that, as a result, they are both fast and similar in speed. However, these two articulators may have other side effects. For example, because they are placed differently in the oral cavity, their speed may have different effects.

Furthermore, the jaw degree opening does not vary between /m/ and /n/ in Swedish speakers (Lindblom, 1983). However, a lower jaw, which is the case with particular nasal coronal consonants, seems to be related not only to greater speaker variability but also to a higher degree of effect from nearby segments (Mooshammer et al., 2006). Furthermore, according to Mooshammer et al. (2007) this pattern is also dependent on the place of articulation - the further back a constriction occurs, the greater the variability. This might indicate a difference between /m/ and /n/ in terms of the effect of nearby segments, as a result of how they are connected to the jaw.

How can any of these differences between /n/ and /m/ explain our results? Surely their effect on the other consonant is what is significant. Actually, regarding our results as regards the place of articulation, the inherent characteristics of the various consonants might not be crucial. Instead, the crucial factor may be whether they are the same. This suggestion is better explained by taking a look again at the collaboration between articulators, the inter-articulatory cohesion, and how it relates to ease of articulation.

Our results can be interpreted as saying that if one has to change the place of articulation within the syllable, this in its turn means an overall increased articulatory complexity. Accordingly, changing the place of articulation could mean a higher demand for coordination, even across the vowel. Thus, coordination complexity might lengthen even non-adjacent segments. How more complex articulation prolongs non-locally is unclear. It could be a matter of coordination, i.e. a phonological difference, which is about an inherent relationship between gestures in the syllable (see e.g. Browman and Goldstein, 1988). Thus, since a speaker might need to activate multiple articulators early, a more complex articulatory map is made, which may take longer time to execute.

The non-local effect as a result of more complex articulation could also be a mainly mechanical influence - that the lips, for example, need to move faster in order to be able to close again in time for the next consonant. The same should apply to the tongue tip; a faster movement in either of C1 and C2 to be able to lower the tongue body sufficiently for the open vowel between the consonants. The results of *variable 2*, which compares the words /ˈamːa/ and /ˈmamːa/, may corroborate the mechanical explanation. The geminate in /ˈamːa/ is significantly longer, which indicates that the lips have taken plenty of time to perform the constriction. Thus, in /ˈamːa/, the lips do not have "competition" and do not suffer from the timing constraint, as the lips in /ˈmamːa/. However, the approximately 30 ms longer C2 in /ˈamːa/ may also include an effect of prominence. In other words, an extended duration may be due to both the absence of C1 and to prominence.

Regardless of the explanatory model, future research should examine whether this phenomenon of shorter segments, because of the same place of articulation, is present in other consonants. Furthermore, the hypothesis as to whether the complexity of changing positions has a non-local effect needs to be articulately tested, to see if the gestures are really different. An ongoing articulatory study indicates that the lips in a word-initial consonant indeed behave differently depending on the subsequent consonant.

5.2 Non-local effect of manner

The results on the manner of articulation were somewhat inconclusive. Although we found a tendency towards an effect of the manner of articulation on C1, these effects were small. In addition, the raw duration results indicated that prominence might have skewed the results. Furthermore, the effect of a different manner of articulation on the word-medial /l/ did not reach significance. Both *variable 3* and *5* investigating manner contained /l/. Laterals are more likely to vary and to be affected by their context; in addition, they are more difficult to segment. Maybe we would have had a clearer result if we had compared the manner of articulation of word-medial C2s that were more consistent, as they are articulated with a higher jaw – such as sibilants, or dorsal consonants. It should also be mentioned that the variables on manner contained word-medial C2 that were post-syllabic, which means that any effect found probably exceeded the syllable boundary. A further discussion of this issue can be found in section 5.4.

In general, the results on short consonants vary less. This applies to C2 in *variables 3* and *5*, and all C1s (Table 2) and may partly explain why we get clearer results as regards the place of articulation, since those comparisons contained long consonants (*variables 1, 2* and partly *4*). Thus, the short consonants do not give much effect because they must remain short for the sake of contrast while the long consonants have the opportunity to vary more without losing their contrast function. In addition, it appears that the non-local effect was generally greater in C2 than in C1, regardless of consonant length. This could be interpreted as CV being more stringent while VC is more variable.

5.3 Non-local effect of geminate

A particularly interesting result was that it seemed that the word /ˈmɑːnar/ had gained more prominence than the other words. It is highly likely that this was also the reason why the /m/ in /ˈmanːa/ was not considerably longer than /m/ in /ˈmɑːnar/. Hence, we did not corroborate Braun/Turco's results (2016). That is, we didn't find a gemination effect on the word-initial /m/. The prominence effect in /ˈmɑːnar/ seems to have overridden a possible lengthening by the geminate.

However, there is also a morphological difference between these two target words – and a more complicated structure in /mɑːn+ar/ than in /manːa/. If morphological complexity were to be part of the explanation, it would mean that a more complex word-medial structure entails word-initial lengthening. We are uncertain if this is a reasonable conclusion. In such a case, this would in some way counteract the gemination effect in that a more complex structure prolonged the segments.

In addition, the comparison of geminate and singleton yields another obvious difference between the target words in *variable 5*. Swedish has compensatory lengthening – a geminate follows a short vowel [a] in a closed syllable; a singleton in a post-syllabic position is preceded by an open syllable with a long vowel [ɑː]. Thus, in Swedish, apart from different syllable structures, the vowel length contrast also signifies vowel quality differences. The vowel quality difference is a factor we were not able to control, and it could be yet another reason why we did not find a gemination effect on the word-initial segment, as was found in Italian (Turco & Braun, 2016). In short, in this study there are too many possible factors that may have affected our results in *variable 6* as regards the gemination effect.

Since we have not produced any clear results on the geminate, we offer no discussion about its articulatory influence. However, the place of articulation between long and short consonants is presumed not to differ. Therefore, the same explanatory model as for the place of articulation is not suitable. Instead, it seems likely that "anticipatory slowness", as suggested in the Introduction, may be the reason of measured C1 differences in other languages. Further cross-linguistic research on geminate lengthening is necessary.

5.4 Post-syllabic non-local effects

The geminate word /ˈnanːa/ was compared with a more complex articulation, and a more complex morphemic structure in the cluster word /namn+ar/ (*variable 7*). Given the results we had already seen regarding the place of articulation, we expected C1 in /ˈnanːa/ to be shorter. Moreover, if our assumptions of morphological complexity and articulatory complexity also hold, it would lead to the same result - that /n/ in /ˈnanːa/ is shorter than in the many ways more complex /ˈnamnar/. However, there was no duration difference in the word-initial /n/. We found instead that the word-initial consonants were of the same length.

What can this be due to? There is a possible explanation model that we haven't touched on yet: the post-syllabic effect. Although it may be inappropriate to discuss the absence of significant difference in the results, we will make an attempt. The reason why we cannot find an effect on C1 in *variable 7* could be because the second slot in the consonant cluster, the post-syllabic onset /n/, influences the preceding /n/ in /ˈnamnar/. In its turn, this could mean that the effect of the place of articulation is not restricted to the syllable. This agrees with previous research that also found non-local effects by post-syllabic segments (Mielke & Nielsen, 2018). It is possible that in a cluster, the second consonant (the onset of the next syllable) is somehow stronger and therefore overrides the first consonant (coda), and its potential relation to the word-initial C1.

As already noted, the short consonants (C1 and C2) seem to vary less than the long consonants (C2) at the same time as word-initial consonants vary less than word-medial ones. However, the short C2s are simultaneously post-syllabic; therefore the fact that there is less variation among the short consonants may be due to syllable position. That is, syllable onset is always more stringent than syllable offset, even if the consonant is in a word-medial position. A closer analysis of /ˈnamnar/ could clarify this, as it is the only short consonant in our material that has a coda position.

Finally, a note on phonological entities: our results suggest that the units in phonological planning are larger than syllables, since we may have found post-syllabic effects. However, we reserve the right to interpret our results less strictly. If an articulatory study were to show that the mechanical effects is a sufficient explanation of the length of the different segments, does that constitute evidence for or against their belonging to the same unit? We need to know more about how phonological relationships are expressed in articulation, and how to interpret them, to be able to answer this. Thus, more research is needed on articulatory phonological structures in order to be able to properly map the meaning-carrying units.

6 Conclusion

In this study, we find a mutual influence of non-local effects through factors (i.e. the place of articulation) that are similar in word-initial and word-medial position. Furthermore, the non-local effect appears to be greater in a word-medial position than in a word-initial position. In addition, the non-local effect seems to affect long consonants more than short consonants. The long consonants are more variable than the short consonants, as the short consonants must preserve the phonological contrast function. However, syllable structure may have affected some of the results - the short consonants may show less variation simply because in Swedish they are post-syllabic in a word-medial position.

We found that the manner of articulation had an effect on the word-medial post-syllabic consonant. However, this result may be due to a prominence that extended certain segments. The prominence phenomenon may also have been the reason why we did not find a gemination effect in the word-initial consonant.

We have discussed various reasons for the results regarding mutual influence of the place of articulation. It seems a reasonable scenario that the more an articulation varies (and therefore can be seen as more dynamically complex) within a word, the more time do the different parts of the mouth need to perform it. Thus, the inherent characteristics of a consonant affect the position that an articulator moves away from, and the distance it travels lengthens or shortens with it. It seems obvious that position is important for a segment's realization, but not just the place of articulation of the segment in question. The position from which an articulator travels, as well as where it is going next, are also significant factors.

7 Acknowledgements

This study is part of a dissertation. The first drafts of the manuscript have been developed in consultation with Professor Martine Grice. I am extremely grateful to her for taking the time to advise me.

I am also grateful to Professor Mattias Heldner, Professor Sven Strömqvist, Dr. Gilbert Ambrazaitis and Dr. Johan Frid for valuable comments on the finished draft of the manuscript.

The author gratefully acknowledges Lund University Humanities Lab.

8 References

Bates, D., Maechler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01

Bell-Berti, F. & Harris, K. (1979). Anticipatory coarticulation: Some implications from a study of lip rounding. *The Journal of the Acoustical Society of America* 65, 1268-1270, doi:10.1121/1.382794

Boersma, P., & Weenink, D. (2018). *Praat: doing phonetics by computer* [Computer software], Version 6.0.37. Retrieved 3 February 2018 from http://www.praat.org/

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology Yearbook*, 3, 219–252.

Browman, C. P., & Goldstein, L. (1988). Some notes on syllable structure in articulatory phonology. *Phonetica*, 45(2–4), 140–155. doi:10.1159/000261823

Bruce, G. (1977). *Swedish word accents in sentence perspective*. Lund: Gleerup.

Bruce, G. (2005). Intonational prominence in varieties of Swedish revisited. In S. Jun (Ed.), *The phonology of intonation and phrasing* (pp. 410–429). Oxford: Oxford University Press.

Cho, T. & Ladefoged, P. (1999). Variation and universals in VOT: evidence from 18 languages. *Journal of Phonetics* 27, 207–229.

Elert, C.-C. (1964). Phonologic studies of quantity in Swedish. Uppsala: Almqvist & Wiksell.

Fowler, C. A., & Saltzman, E. (1993). Coordination and coarticulation in speech production. *Language and Speech*, 36(2–3), 171–195. doi:10.1177/002383099303600304

Gårding, E. & Lindblad, P. (1973). Constancy and variation in Swedish word accent patterns. *Working Papers, Phonetics laboratory, Lund University* 7, 36–110

Harrington, J. (2010). Acoustic Phonetics. In *The handbook of phonetic sciences*/edited by William J. Hardcastle, John Laver, Fiona E. Gibbon. – 2nd ed, 81–129.

Heldner, M. (2001). On the non-linear lengthening of focally accented Swedish words. In W. van Dommelen & T. Fretheim (Eds.), *Nordic Prosody: Proceedings of the VIIIth Conference, Trondheim 2000* (pp. 103-112). Peter Lang.

Iskarous, K., Fowler, C. and Whalen, D. (2010). Locus equations are an acoustic expression of articulator synergy. *The Journal of the Acoustical Society of America* 128, 2021; https://doi.org/10.1121/1.3479538

Kawahara, S., Masuda, H., Erickson, D., Moore, J., Suemitsu, A., Shibuya, Y. (2014). Quantifying the effects of vowel quality and preceding consonants on jaw displacement: Japanese data. *Journal of Phonetic Society of Japan*, 132, 54–62.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, 82(13), 1–26. doi:10.18637/jss.v082.i13

Lindblom, B. (1983). Economy of speech gestures. In MacNeilage, P. (ed) *The Production of Speech.* New York: Springer-Verlag. 217–245.

Löfqvist, A. (2005). Lip kinematics in long and short stop and fricative consonants. *The Journal of the Acoustical Society of America* 117, 858; doi:10.1121/1.1840531

Löfqvist, A. (2006). Interarticulator programming: Effects of closure duration on lip and tongue coordination in Japanese. *The Journal of the Acoustical Society of America* 120, 2872; doi: 10.1121/1.2345832

Löfqvist, A. (2007). Tongue movement kinematics in long and short Japanese consonants, *The Journal of the Acoustical Society of America* 122 (1), 512–518.

Mielke, J. & Nielsen, K. (2018). Voice Onset Time in English voiceless stops is affected by following postvocalic liquids and voiceless onsets. *The Journal of the Acoustical Society of America* 144, 2166; doi:10.1121/1.5059493

Mooshammer, C., Hoole, P. & Geumann, A. (2006). Interarticulator cohesion within coronal consonant production. *The Journal of the Acoustical Society of America,* 120(2), 1028–1039. doi:10.1121/1.2208430

Mooshammer, C., Hoole, P. & Geumann, A. (2007). Jaw and order. *Language and Speech,* 50(2), 145–176

R Core Team (2015). *R: A language and environment for statistical computing* [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Rosen, K. (2005). Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison, *Journal of Phonetics* 33, 411–426. doi:10.1016/j.wocn.2005.02.001

Smith, C. (1995). Prosodic patterns in the coordination of vowel and consonant gestures. In B. Connell, & A. Arvaniti (Eds.), *Laboratory Phonology IV: Phonology and Phonetic Evidence* (pp. 205–222). Cambridge: Cambridge University Press.

Svensson Lundmark, M., & Frid, J. (2019). Jaw movements in two tonal contexts. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, 1843–1847.

Svensson Lundmark, M., Frid, J., Ambrazaitis, G. & Schötz, S. (submitted) Word-initial CV coarticulation in a pitch-accent language.

Turco, G. & Braun, B. (2016). An acoustic study on non-local anticipatory effects of Italian length contrast. *The Journal of the Acoustical Society of America*, 140, 2247–2256. doi:10.1121/1.4962982

Türk, H., Lippus, P. & Simko, J. (2017). Context-dependent articulation of consonant gemination in Estonian. *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 8(1), 1–26.

Wieling, M., & Tiede, M. (2017). Quantitative identification of dialect-specific articulatory settings. *The Journal of the Acoustical Society of America*, 142(1), 389–394. doi:10.1121/1.4990951

Xu, Y. (2013). ProsodyPro — a tool for large-scale systematic prosody analysis. In B. Bigi, & D. Hirst (Eds.), *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France 2013 (pp. 7–10). Aix-en-Provence, France: Laboratoire Parole et Langage.

Öhman, S. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, 39(1), 151–168. doi:10.1121/1.1909864

Öhman, S. (1967). Word and sentence intonation: A quantitative model, *Quarterly Progress and Status Report, Dept. for Speech, Music and Hearing, KTH Stockholm*, 8(2-3), 20–54.